Détection de motifs audio pour la séparation de sources guidée. Application aux bandes- son de films.

Nathan Souviraà-Labastie

ANNÉE 2015

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Traitement du signal et télécommunications

École doctorale Matisse

présentée par

Nathan SOUVIRAÀ-LABASTIE

préparée à l'unité de recherche IRISA – UMR6074
Institut de Recherche en Informatique et Système Aléatoires
Université de Rennes 1

Détection de motifs audio pour la séparation de sources guidée. Application aux bandes-son de films.

Thèse soutenue à Rennes le 23 Novembre 2015

devant le jury composé de :

Bertrand DAVID
Maître de conférences, Télécom ParisTech / *Rapporteur*

Christian JUTTEN
Professeur, Université Joseph Fourier / *Rapporteur*

Régine ANDRÉ-OBRECHT
Professeur, Université de Toulouse / *Examinatrice*

Christian UHLE
Senior Scientist, Fraunhofer IIS / *Examinateur*

Frédéric BIMBOT
Directeur de recherche, CNRS-IRISA / *Directeur de thèse*

Emmanuel VINCENT
Chargé de recherche, Inria Nancy-Grand Est / *Co-directeur de thèse*

Remerciements

Par ordre chronologique,

je tiens à remercier en premier lieu mon directeur de thèse Frédéric sans qui je n'aurais pas pris goût à la recherche. Merci d'avoir ensuite réuni tous les ingrédients me permettant de me lancer dans cette aventure au long cours. Merci notamment d'avoir pensé à Emmanuel pour co-encadrer cette thèse. Merci à toi Emmanuel pour ta patience et ton encadrement. Ce manuscrit n'aurait pas cette finition et cette justesse sans ta présence. Merci à vous deux pour votre complémentarité tant sur le plan scientifique qu'humain (et organisationnel aussi). Je remercie également Christophe Henrotte et l'ensemble du Studio Maia pour l'expertise d'ingénieur du son apportée à cette thèse.

Merci à tous les anciens de METISS, aux nouveaux de PANAMA, aux collègues de projet, de rédaction, aux coincheurs, aux débatteurs du Supélec vs Biocoop vs RU et autres débatteurs de mauvaise foi. Merci pour tout ce café et toutes ces bières. Merci à toi, Laurence, d'apparaître dans toutes ces catégories et d'autres encore. Merci à toi, Jules, co-bureau d'exception, pour l'ensemble de ton œuvre. Je remercie aussi l'équipe PAROLE/MULTISPEECH et ses membres pour leur accueil durant la période la plus intensive de cette thèse. Merci aussi à toi Antoine pour toutes ces idées, explorées et inexplorées.

Je tiens à remercier l'ensemble des membres du jury pour leur intérêt et leur bienveillance. Je remercie en particulier mes rapporteurs Bertrand David et Christian Jutten pour leur relecture fine et complémentaire. Sans compter les améliorations directes du manuscrit, vos commentaires éclairés m'ont ouvert de nouvelles perspectives. Je remercie également Christian Uhle pour son enthousiasme et sa présence, ainsi que Mme André-Obrecht pour avoir présidé ce jury.

Enfin, en dehors de tout ordre chronologique, je te remercie toi Emilie, d'avoir été là sans discontinuer et de m'avoir supporté.

Résumé

Lorsque l'on manipule un signal audio, il est généralement utile d'opérer un isolement du ou des éléments sonores que l'on cherche à traiter. Cette étape est couramment appelée **séparation de sources audio**. Il existe de nombreuses techniques pour estimer ces sources et plus on prend en compte d'informations à leur sujet plus la séparation a des chances d'être réussie. Une façon d'incorporer des informations sur une source est l'utilisation d'un signal de référence qui va donner une première approximation de cette source. Cette thèse s'attache à explorer les aspects théoriques et appliqués de la **séparation de sources audio guidée par signal de référence**. La nouvelle approche proposée appelée *SPotted REference based Separation* (SPORES) examine le cas particulier où les références sont obtenues automatiquement par **détection de motif**, c'est-à-dire par une recherche de contenu similaire. Pour qu'une telle approche soit utile, le contenu traité doit comporter une certaine **redondance** ou bien une large base de données doit être disponible. Heureusement, le contexte actuel nous permet bien souvent d'être dans une des deux situations et ainsi de retrouver ailleurs des motifs similaires. L'objectif premier de ce travail est de fournir un cadre théorique large qui une fois établi facilitera la mise au point efficace d'outils de traitement de contenus audio variés. Le second objectif est l'utilisation spécifique de cette approche au traitement de **bandes-son de films** avec par exemple comme application leur conversion en format surround 5.1 adapté aux systèmes *home cinema*.

Abstract

In audio signal processing, ***source separation*** *consists in recovering the different audio sources that compose a given observed audio mixture. They are many techniques to estimate these sources and the more information are taken into account about them the more the separation is likely to be successful. One way to incorporate information on sources is the use of a reference signal which will give a first approximation of this source. This thesis aims to explore the theoretical and applied aspects of* ***reference guided source separation****. The proposed approach called SPotted REference based Separation (SPORES) explore the particular case where the references are obtained automatically by* ***motif spotting****, i.e., by a search of similar content. Such an approach is useful for contents with a certain* ***redundancy*** *or if a large database is be available. Fortunately, the current context often puts us in one of these two situations and finding elsewhere similar motifs is possible. The primary objective of this study is to provide a broad theoretical framework that once established will facilitate the efficient development of processing tools for various audio content. The second objective is the specific use of this approach to the processing of* ***movie soundtracks*** *with application in 5.1 upmixing for instance.*

Table des matières

Liste des figures

Liste des tableaux

Liste des acronymes

CQT *Contant Q Transform.*

DEMIX *Direction Estimation of Mixing matrIX.*
DNN *Deep Neural Network.*
DTW *Dynamic Time Warping.*
DUET *Degenerate Unmixing Estimation Technique.*

EM *Expectation-Maximization.*

FFT *Fast Fourier Transform.*

GCC *Generalized Cross Correlation.*
GEM *Generalized Expectation-Maximization.*
GMM *Gaussian Mixture Model.*

HMM *Hidden Markov Model.*

ICA *Independent Component Analysis.*
IS Itakura-Saito.
ISTFT *Inverse Short Time Fourier Transform.*

KL Kullback-Leibler.
KWS *Keyword Spotting.*

MFCC *Mel-Frequency Cesptral Coefficient.*
MIDI *Musical Instrument Digital Interface.*
MMSE *Minimum Mean Square Error.*
MU *Multiplicative Updates.*

NMF *Nonnegative Matrix Factorization.*
NMPcF *Nonnegative Matrix Partial co-Factorization.*

PHAT *PHAse Transform.*

PLCA *Probabilistic Latent Component Analysis*.

REPET *REpeatting Pattern Extraction Technique*.

SAOC *Spatial Audio Object Coding*.
SAR *Signal-to-Artifacts Ratio*.
SCA *Sparse Component Analysis*.
SDR *Signal-to-Distortion Ratio*.
SIR *Signal-to-Interference Ratio*.
SNR *Signal-to-Noise Ratio*.
SPORES *SPotted REference based Separation*.
STFT *Short Time Fourier Transform*.

TDOA *Time Delay Of Arrival*.

Chapitre 1

Introduction

Le traitement du signal audio est un vaste domaine scientifique proche d'autres domaines également en expansion comme le traitement des images, l'apprentissage automatique (*Machine Learning*) ou le Traitement Automatique des Langues (TAL). De nombreux types de traitements ou analyses peuvent être appliqués au signal audio dans sa grande variété de contenus : cinéma, télé, radio... Cependant l'oreille humaine est d'une rare sensibilité. Que ce soit pour localiser, isoler ou identifier un son, le système auditif humain est doté de capacité d'analyse importante. Le traitement automatique et la production de contenus audio doivent donc répondre à certaines exigences.

Dans le but d'augmenter les possibilités de traitement et/ou de production audio, il est utile de pouvoir opérer un isolement du ou des éléments sonores que l'on cherche à traiter. Cette étape est couramment appelée **séparation de sources audio**, mais aussi dissociation sonore. Les contenus concernés sont divers (films, musique polyphonique, parole bruitée...) de même que les techniques utilisées (aveugles, guidées, informées) [111, 173].

Cette thèse s'attache à explorer une nouvelle approche de séparation de sources qui exploite la **redondance** des signaux. Je propose pour cela de rapprocher des techniques de détection de motifs et de séparation de source guidée par signal de référence. Après avoir placé ces techniques dans leur contexte scientifique et applicatif, ce chapitre introductif décrit les verrous et défis qu'implique cette nouvelle approche.

1.1 Séparation de sources

La présence de bruit et plus généralement la superposition de sources audio (par exemple les voix de plusieurs locuteurs ou plusieurs instruments de musique) peut perturber les traitements du signal audio [171, 179]. La séparation de sources audio [1] est la tâche qui consiste à isoler les sources audio les unes des autres, et est une possibilité pour améliorer ces traitements.

1.1.1 Notion de mélange

Le signal audio peut contenir des éléments assez variés comme de la voix, de la musique ou encore des sons d'ambiance. Quand plusieurs éléments ou sources sont présents, on parle alors de mélange. Le mélange est dit instantané lorsqu'on le considère comme la simple addition de ses sources. En revanche, on parle de mélange convolutif si on prend en considération la propagation ou la réverbération.

Pour ce qui est de la localisation des sources, plusieurs canaux audio sont nécessaires pour en rendre compte. Un enregistrement qui regroupe ces différents canaux est alors dit **multicanal**. Le mélange est dit sur-déterminé si il y a plus de canaux que de sources et sous-déterminé si il y a plus de sources que de canaux. Nous verrons plus tard que cela influence le choix des techniques de séparation.

Un tel signal multicanal peut résulter d'un réel enregistrement physique (plusieurs microphones) ou d'une spatialisation artificielle des sources comme dans le cas d'un mixage stéréo en studio. Différents modèles permettent alors de décrire l'emplacement d'émission d'une source. L'hypothèse la plus utilisée mais souvent incomplète est de considérer que le son est émis à partir d'un seul point de l'espace (source ponctuelle). À l'inverse, on parle de source diffuse lorsqu'on considère une zone plus large d'émission [45], par exemple si une caisse de résonance est impliquée.

1.1.2 Séparation de sources aveugle ou guidée

Historiquement, le problème de la séparation sur-déterminée est traité par des approches aveugles, c'est-à-dire qui n'utilisent pas d'information à priori sur les sources. En effet, il a été prouvé que dans ce cas l'estimation du processus de mélange est équivaut à l'estimation des sources [18]. Dans le cas sous-déterminé, l'estimation du processus de mélange n'est pas suffisante pour effectuer la séparation et la modélisation des sources est requise [18]. On va en contrepartie s'attacher à estimer les fréquences et les instants auxquels chaque source est active. Les méthodes dites aveugles [37] sont donc limitées dans le cas sous-déterminé et les approches guidées sont à privilégier, en particulier pour les enregistrements mono-canal [1] et pour les applications professionnelles.

Les méthodes de **séparation de sources guidée** regroupent un certain nombre de techniques qui permettent de prendre en compte des informations à propos des sources [111, 173]. Différents types d'informations extérieures peuvent être incorporés, par exemple des informations symboliques sous forme de fichier MIDI (*Musical Instrument Digital Interface*) [60] ou de texte [105]), des informations données par un utilisateur [44, 131, 157] ou encore sous forme de signal [113, 157]. Nous nous intéresserons particulièrement à ce dernier cas aussi appelé séparation guidée par signal de référence ou encore *exemplar-based separation* comme défini par Liutkus *et al.* [111]. Ces approches guidées sont à ne pas confondre avec la séparation de sources informée [110, 137] dont le but est le codage des sources et leur transmission.

1. Lorsque le mélange est mono-canal, le problème de la localisation ne peut pas être résolu en général et les sources ne sont pas identifiables selon leurs angles d'arrivée.

1.2 Contexte applicatif

Les applications de la séparation de sources peuvent être classés en deux catégories, celles qui relèvent principalement de l'analyse (reconnaissance, transcription) des contenus et celles qui visent à modifier ces contenus. Dans le cas de systèmes d'analyse, la séparation de sources peut être utilisée comme prétraitement pour rendre ces systèmes plus robustes, par exemple pour la reconnaissance de la parole en environnement bruité. Les différentes sources obtenues par la séparation peuvent aussi faire l'objet d'une réutilisation comme par exemple pour le post-traitement de bandes-son de films qui est l'application finale visée.

En ce qui concerne la détection de motifs, l'application principale est l'archivage des contenus audio ce qui permet de parcourir ou encore de structurer le contenu.

1.2.1 Le post-traitement des bandes-son de films

L'arrivée sur le marché de nouveaux supports de distribution (DVD puis Blu-Ray) a fait évoluer les standards de qualité, notamment en rendant possible la diffusion des bandes-son dans différentes langues et en format *surround*[2] (5.1 et plus) adapté aux systèmes *home cinema* (Figure 1.1). Ces évolutions impliquent des changements lors de la phase de montage des bandes-son de nouveaux films et créent un nouveau marché en ce qui concerne des films plus anciens que l'on veut distribuer au standard de qualité actuel (rehaussement des couleurs ou colorisation, restauration de la bande-son et passage en format *surround*).

Figure 1.1 – Système audio de diffusion du format surround 5.1 pour *home cinema*.

Le processus de conversion d'une bande-son en format *surround* est appelé ***upmixing***. On cherche alors à spatialiser les sources sonores de façon cohérente avec les actions du film ou en suivant les conventions cinématographiques. Disposer des pistes séparées pour chaque élément sonore est alors indispensable pour réaliser cette conversion. Lorsque les pistes ne sont pas disponibles (notamment pour une seconde post-production) il est nécessaire d'opérer la séparation des sources. Les sources à dissocier sont dans ce cas les voix et dialogues, les bruitages, la musique, l'ambiance et les effets sonores.

De la même façon que pour l'*upmixing*, la génération d'une **version internationale** (version sans les dialogues) en seconde post-production fait aussi appel à la séparation de sources audio. Ce besoin de séparation apparait par exemple lorsque les pistes n'ont pas été sauvegardées à la suite d'une précédente post-production.

La réutilisation des sources audio (*repurposing*) [8, 61] concerne aussi des applications en musique. L'*upmixing* de morceaux de musique, la génération automatique de

2. Les formats dit *surround* incluent des canaux pour les haut-parleurs situés derrière l'auditeur, lui donnant ainsi l'impression que les éléments sonores l'entourent.

karaoké [75, 162] (retrait du chanteur) ou de *backing track* (isolement d'un instrument) sont respectivement très similaires à l'*upmixing* et à la génération de versions internationales de bandes-son de films. La re-égalisation d'instrument ou le *sampling* d'extraits musicaux sont aussi possibles que ce soit pour un professionnel (*remastering*, création) ou pour le grand public (*active listening*).

Dans tous les cas, une séparation n'est jamais parfaite et induit des effets indésirables qu'il convient de prendre en compte ou de traiter à posteriori. Pour les applications où les sources sont réutilisées, une phase manuelle de nettoyage est obligatoire pour obtenir une qualité de niveau studio. Si les sources sont réutilisées pour créer un nouveau mélange, les artefacts de la séparation pourront être cachés par les nouveaux éléments, notamment si on réutilise des sources provenant de la même séparation (par exemple l'égalisation d'instrument).

Cette thèse s'inscrit en outre dans le cadre d'une **collaboration industrielle** avec le Studio Maia. Ce studio offre des services de post-production de bandes-son de films qui nécessitent au préalable la séparation des sources audio. Le principal objectif est l'automatisation partielle du processus de séparation. L'ingénieur du son devra cependant rester au centre du dispositif de post-production pour garantir une très bonne qualité finale. Les traitements que je propose doivent donc pouvoir s'intégrer dans un processus semi-automatique. Une description plus étendue de cette collaboration est donnée dans le chapitre 8.

1.2.2 Reconnaissance et transcription

Pour l'Homme, le signal sonore n'est souvent que le support qui permet de transporter de l'information (alerte, langage, musique, émotion...). Le contenu suit généralement des règles ou des structures préétablies (« encodage ») : les phonèmes, les mots et les phrases pour la parole ; le rythme, les accords et les différents instruments pour la musique. La reconnaissance de tels « encodages » nécessite donc une compréhension de ces structures à plusieurs niveaux. D'une façon générale, les techniques de reconnaissance du signal peuvent être décrites comme des systèmes de conversion du signal en éléments symboliques.

De nombreuses applications et produits commerciaux découlent de toutes les techniques répertoriées ici. On peut rapidement citer l'apprentissage pédagogique des langues ou de la musique ou encore les logiciels de dictée vocale et de commande vocale qui sont commercialisés depuis quelques années, par exemple Siri de Apple®.

Reconnaissance de la parole En ce qui concerne la reconnaissance de la parole, la première étape est de reconnaitre des unités acoustiques, appelés phones. Le signal audio est converti à intervalles de temps réguliers en vecteurs de descripteurs, typiquement des coefficients cepstraux en échelle Mel (*Mel-Frequency Cesptral Coefficient* ou MFCC) [55]. La mise en correspondance sur chaque intervalle/trame temporelle des descripteurs et des unités acoustiques était historiquement réalisée par des modèles de mélange de gaussiennes (*Gaussian Mixture Model* ou GMM). Mais l'évolution récente des techniques d'apprentissage automatique [92] a permis aux réseaux de neurones pro-

fonds (*Deep Neural Network* ou DNN) de surpasser cette technique pour estimer la probabilité de chaque unité acoustique [91]. Les probabilités estimées sont ensuite exploitées à plus haut niveau pour déterminer quel est le mot ou la phrase la plus probable, grâce à un modèle de langage et un lexique. L'approche la plus usuelle repose alors sur les modèles de Markov cachés (*Hidden Markov Model* ou HMM) [143]. On peut aussi signaler l'existence de techniques de reconnaissance du locuteur ou de la langue.

Dans le cas de la reconnaissance de la parole en environnement réel [171,182], on peut considérer que le signal est composé d'une source de parole et d'une ou plusieurs sources de bruit. La séparation de sources peut alors fournir une version débruitée du signal de parole à analyser. Des mesures de confiance (incertitudes) sur les descripteurs peuvent alors être calculées et prises en compte par un système de reconnaissance robuste de la parole [165–167] ou du locuteur [147].

Transcription de la musique En musique, on parle de transcription [16, 102], le terme « reconnaissance » étant associé à la tâche d'identification de morceaux. La **transcription symbolique de la musique** regroupe différents scénarios : la transcription monophonique (un instrument jouant une note à la fois), et la transcription polyphonique (un instrument jouant plusieurs notes ou plusieurs instruments). Dans chaque cas, on cherche à estimer les paramètres des notes jouées (symboles), principalement la hauteur, la durée et le timbre, mais aussi l'intensité. Les résultats peuvent par exemple être regroupés dans un fichier au format MIDI. Des techniques spécifiques existent pour l'estimation de ces différents paramètres comme l'estimation de la hauteur, l'estimation du tempo ou encore la reconnaissance des instruments. Chaque technique peut bénéficier des autres, mais résoudre le problème dans sa totalité est une tâche complexe. De plus, les interconnexions sont rendues difficiles par les natures diverses des paramètres à estimer. Aucune approche générale ne s'est pour le moment imposée dans la communauté à l'instar du couple DNN/HMM [91] pour la reconnaissance de la parole qui est un problème davantage standardisé.

La séparation de sources peut être bénéfique pour certaines sous-tâches de la transcription [16] comme l'identification des instruments [23, 94], l'estimation de hauteurs multiples [56,156,169], ou encore l'extraction de descripteurs [10]. De même que pour le traitement de la parole, la propagation de l'incertitude peut aussi présenter un intérêt comme par exemple pour la reconnaissance du chanteur [104].

1.2.3 Archivage et détection de motif

L'**archivage** est « l'action de conserver et de classer des documents ne présentant plus un intérêt immédiat » [7]. Sans un classement approprié une grande collection de documents devient difficile d'usage voire inutilisable notamment après dématérialisation [148]. On peut par exemple classer des documents par leur auteur, ou regrouper leurs thèmes dans un index. Par exemple, l'index d'un livre est une liste de mots-clefs accompagnés de leurs références (dans ce cas les numéros de pages où le mot-clef apparait). On peut ainsi gagner un temps précieux en parcourant l'index plutôt que tout le contenu à la recherche d'un mot. De même il peut être utile de créer le même genre

d'index de mots-clefs pour un document audio. La création automatique des références (instants où apparait chaque mot) d'un tel index requiert alors l'utilisation d'un algorithme de détection de motifs audio adapté à la parole.

De façon plus générale, la **détection de motifs audio** est la tâche qui consiste à détecter/rechercher dans une base de données tous les motifs audio qui ressemblent à une requête d'entrée. Cette requête est un signal audio qui peut être un mot, un passage musical, ou encore une chanson entière. Les motifs recherchés peuvent être dans une certaine mesure déformés, par exemple un mot prononcé par une autre personne, un morceau de musique repris par d'autres musiciens ou encore une version bruitée de la requête. Les représentations du signal audio utilisées pour effectuer ces comparaisons doivent donc être invariantes, c'est-à-dire tolérantes, à ces déformations.

Lorsque les déformations entre la requête et le motif recherché sont faibles on parle alors d'**identification audio** [63, 81, 179]. Ces déformations sont typiquement introduites durant la transmission d'un des signaux (réseau téléphonique ou microphone de basse qualité). Dans ce cas, on peut utiliser des représentations succinctes du signal audio. En revanche, on sera intéressé par l'identification rapide de la requête au sein d'une vaste base de données. La difficulté réside alors en grande partie dans la phase d'**indexation** [63, 81, 141] qui permet d'accélérer la tâche, la table d'index étant grande. On cherche donc à optimiser le parcours de l'index, à la façon du classement par ordre alphabétique des mots dans un index. Pour prolonger la métaphore, il s'agit en fait de définir le meilleur alphabet pour que le parcours de l'index soit optimisé. Le service d'identification de musique Shazam® est l'exemple type d'application de ces procédés [179].

Alors que les représentations succinctes utilisées pour l'identification résument chaque partie du contenu, il peut aussi être intéressant de résumer tout le contenu [11, 19, 77]. On peut pour cela utiliser un algorithme de **découverte de motifs audio** [28, 120, 123] qui est la variante non supervisée de la détection de motifs : il n'y a pas de requête d'entrée et toutes les parties du document sont comparées les unes avec les autres. Ainsi, on créera un index avec tous les motifs qui se répètent plus de deux fois. Cependant, l'approche la plus usuelle est de résumer le contenu au niveau symbolique. On utilise pour cela les transcriptions des contenus audio (texte, partition) qui peuvent être obtenues par reconnaissance automatique.

1.3 Approche SPORES

Après ce tour d'horizon, on peut retenir que les techniques de séparation et de détection envisagées dans cette thèse ont des environnements déjà riches et imbriqués à d'autres tâches comme la reconnaissance ou l'archivage. On peut aussi entrevoir que de nouvelles synergies sont possibles. C'est ce que cette thèse cherche à faire en proposant de tirer partie de la redondance des signaux pour leur séparation. Je chercherai en particulier à utiliser des techniques de détection de motifs pour identifier les redondances.

1.3.1 Exploitation de la redondance pour la séparation

L'automatisation de la production des contenus, et de façon générale l'augmentation de la quantité de données favorisent l'apparition de contenus redondants : reprise de musique, bases de données sonores publiques... De même, la répétition est une forme de création artistique, en particulier pour les œuvres musicales (styles musicaux répétitifs, création assistée) et pour les œuvres audiovisuelles (cinéma, séries, reportage...). De plus la prise en compte d'informations supplémentaires pendant la séparation permet d'accéder à une meilleure qualité de séparation et ouvre la porte à plus d'applications. L'exploitation de la redondance des contenus audio pour leur séparation apparait donc comme une approche naturelle et prometteuse.

L'approche désignée par *SPotted REference based Separation* (SPORES) sur laquelle porte cette thèse est une possibilité d'approche de séparation de sources qui tire parti de la redondance des contenus (voir Figure 1.2). Il en existe d'autres comme par exemple *REpeatting Pattern Extraction Technique* (REPET) qui est décrite dans la partie 4.2.

On considère que la redondance utilisée pour guider la séparation se situe soit ailleurs dans le même contenu (la même chanson, le même film) soit dans un autre contenu ayant un lien avec le segment à séparer (discographie d'un artiste, bande originale de film, base de données d'effets sonores...). Dans les deux cas, la redondance fait partie d'un grand ensemble et n'est pas encore identifiée. L'approche SPORES prend en entrée un mélange à séparer et un contenu de plus grande dimension susceptible de contenir des éléments similaires.

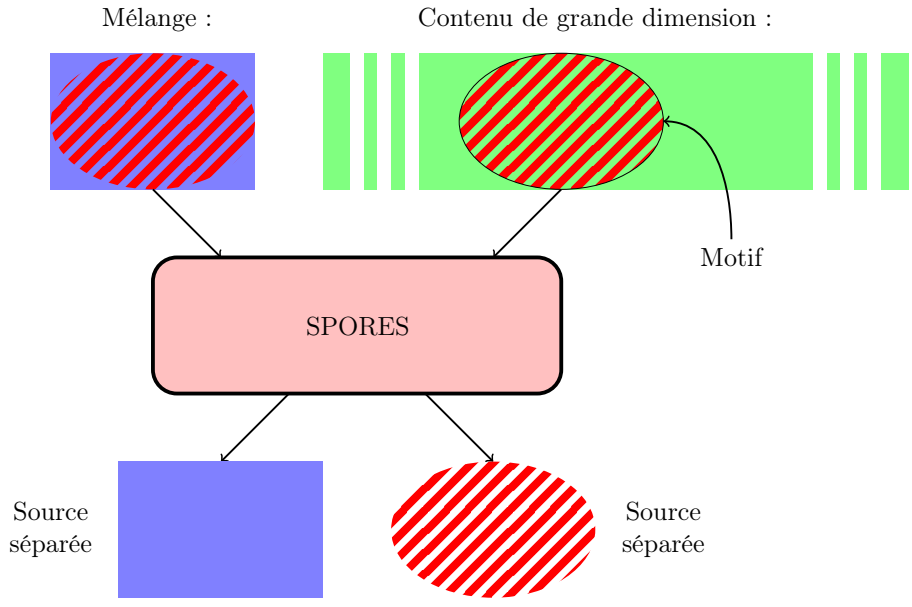


Figure 1.2 – Diagramme de l'approche SPORES.

Sous-tâches de traitement On identifie deux étapes :

— Le mélange sert de requête pour détecter dans ce contenu de grande dimension un motif similaire à une des sources présente dans le mélange. La redondance est typiquement identifiée par détection de motifs.

— Les motifs retrouvés sont ensuite utilisés comme signaux de référence pour guider la séparation de sources. Les sorties de l'approche sont les sources estimées par cette seconde étape de séparation.

Ces deux sous-tâches sont décrites ci-après.

1.3.2 Obtention des références par détection robuste de motifs

La détection de motifs est l'étape de l'approche SPORES qui fournit les références. Dans le cadre de l'approche SPORES les requêtes sont la plupart du temps des mélanges et non des sources isolées et les répétitions des sources ne sont pas toujours exactes. Cette étape doit donc présenter deux facteurs de robustesse qui sont :

— la robustesse aux déformations entre la requête et les motifs recherchés

— et la robustesse à la présence de plusieurs sources dans la requête et dans le contenu de la recherche.

Les techniques précédemment présentées comportent ces deux facteurs de robustesse mais pas conjointement. Les techniques de recherche de mots-clefs sont robustes aux déformations telles que le changement de locuteur mais peu robustes au bruit. À l'inverse, les techniques d'identification de musique sont robustes à certains types de bruits relativement stationnaires [179] mais pas aux déformations telles que le changement d'instrument.

On peut noter que dans un contexte supervisé (présence d'un utilisateur), l'utilisation d'une approche non robuste au bruit est envisageable par exemple si l'utilisateur enregistre la requête sans autre source, ou bien si il sélectionne une zone où il sait qu'il n'y a qu'une seule source.

1.3.3 Séparation de sources guidée par signal de référence

La séparation de sources guidée par signal de référence est une catégorie d'approches de séparation qui utilisent en plus du mélange des informations supplémentaires contenues dans un signal appelé référence.

Une première hypothèse est que la référence comporte une source qui a des propriétés communes avec une des sources présentes dans le mélange. La source et sa référence peuvent par exemple avoir été produites par le même locuteur ou instrument, ou contenir la même séquence de phonèmes ou de notes. Pour que cette approche ait un intérêt, la propriété commune aux deux sources doit être plus facilement observable dans la référence. En d'autres termes, il est préférable que la référence ne soit pas un mélange. On cherche alors principalement à modéliser la déformation entre la source et sa référence.

Une seconde situation est celle où le signal source présent dans la référence et le signal source que l'on cherche à estimer sont les mêmes [109, 113]. Le problème peut alors revenir à l'estimation des différents processus de mélange (celui du mélange à

séparer et celui de la référence). En effet, la référence est bien un mélange dans ce cas. Dans le cas contraire, on détiendrait la vraie source et il n'y aurait plus besoin de séparation.

Les expériences présentées dans ce manuscrit se placent toutes dans le cas sous-déterminé, si on suit strictement la définition précédemment énoncée. Cependant dans le cas de la séparation de sources guidée par signal de référence, il est difficile d'appliquer cette terminologie car dans la plupart des situations on utilise d'autres signaux (en plus des canaux du mélange).

1.4 Évaluations et métriques

L'élaboration de toutes ces techniques d'analyse et de traitement du signal audio passe par une phase d'évaluation et l'utilisation de métriques appropriées à chaque tâche. Le meilleur scénario pour évaluer un système sur une tâche est de le placer dans un cadre contrôlé, c'est-à-dire de connaitre les sorties idéales (**vérité terrain**) pour un jeu de données d'entrée. On utilise alors la ou les métriques adaptées pour comparer les sorties du système à la vérité terrain. Dans des scénarios en temps réel ou pour des systèmes traitant de grands jeux de données, on cherchera en plus à évaluer le **temps de réponse** du système.

Les mesures de **distorsion** entre deux signaux seront elles utiles pour évaluer un encodage ou une qualité de séparation. La mesure de la qualité de séparation [174] peut être complétée par les niveaux d'**interférence** entre les sources estimées, ainsi que des **artefacts** introduits. Des mesures **psychoacoustiques** [57,58] existent également pour ces deux tâches. L'évaluation par des humains (mesures **subjectives**) peut également donner des résultats plus à même de mesurer le potentiel d'un système à visée commerciale. Largement utilisée pour les encodeurs audio (et vidéo), cette méthode deviendra sans doute la norme pour les tâches de séparation dans les années à venir.

Les métriques d'évaluation symboliques permettent de mesurer la ressemblance entre deux séquences de symboles et donc d'évaluer les systèmes de transcription, sachant que les erreurs symboliques peuvent être de plusieurs types (substitution, insertion, omission). Les techniques de reconnaissance sont généralement comparées en terme de taux d'erreur global.

Pour le cas de la détection de motifs, on utilise la précision qui est le taux de bonne réponse du système, qui est lié au nombre d'insertions. Cependant on doit aussi mesurer la capacité du système à retrouver l'ensemble des occurrences, leur nombre lui étant inconnu. Cette aptitude est mesurée par le **rappel** qui est le rapport entre le nombre d'occurrences bien détectées et le nombre d'occurrences à détecter. Ce rapport est lié au nombre d'omissions. L'évaluation de la découverte de motifs est évaluée par des moyennes de précisions et de rappels ou par des variantes de ces métriques [159].

1.5 Défis

Le rapprochement proposé des techniques de détection de motifs et de séparation de sources nous met face à de nouveaux défis.

Premièrement, les techniques existantes de détection de motifs doivent être adaptées au cas des mélanges. Cela implique une robustesse à l'ajout de sources tout en conservant les avantages actuels de robustesse à la déformation des motifs et de temps de calcul réduit.

Ensuite, il est nécessaire de modéliser la déformation entre la source et sa référence durant l'étape de séparation. Certaines approches de séparation permettent déjà de traiter certains cas, mais les déformations rencontrées sont de natures et d'intensités beaucoup plus diverses. Ainsi, il est possible que la référence qui soit trouvée par la détection de motifs ne soit pas exploitable par les approches existantes de séparation. Il est donc indispensable de développer une approche générale de séparation qui permette de prendre en compte n'importe quel type de référence, si elle a pu être caractérisée au préalable.

L'objectif applicatif de ce travail est la mise en place d'outils semi-automatiques pour la séparation de bandes-son de films. Ce contexte nous impose de développer des outils rapides et flexibles d'utilisation. En contrepartie, il a l'avantage de la présence d'un utilisateur expert (ingénieur du son) dans la chaîne de traitement. En effet, l'utilisateur est capable de séparer manuellement des sources et peut donc nous fournir des informations très précises mais qu'il faut utiliser avec parcimonie pour parvenir à une automatisation la plus importante possible.

Un autre défi plus général est de développer deux familles d'outils (détection et séparation) qui restent indépendants et flexibles pour pouvoir être utilisés dans d'autres situations.

1.6 Contributions

Détection de motif Une méthode de détection de motifs déformés robuste à la présence d'autres sources est présentée et validée expérimentalement sur des données artificielles.

Modèle général de déformations Un modèle général de déformation pour la séparation guidée par signal de référence est proposé et expérimenté sur plusieurs types de signaux mono-canal avec estimation des paramètres par l'algorithme de l'état de l'art *Nonnegative Matrix Partial co-Factorization* (NMPcF) [65, 105, 106] :

- modélisation d'un changement artificiel de hauteur [161],
- séparation de mélanges artificiels voix/musique guidée par des références de voix (les mêmes phrases énoncées par d'autres locuteurs) et de musique (motifs obtenus par détection de motifs déformés) [160, 161],
- séparation d'instruments dans un morceau de musique original guidée par les différentes pistes d'une reprise [162].

Ces travaux établissent également une nomenclature des problèmes de séparation guidée par signal de référence.

Algorithmes de séparation multicanale de type *Expectation-Maximization* (EM)

- un algorithme d'estimation multi-mélange/multicanale pour l'estimation des paramètres du modèle général de déformation dans le cas de mélanges multicanaux et sa validation expérimentale pour la séparation de morceaux de musique stéréo guidée par les pistes d'une reprise [161],
- un algorithme adapté aux références très faiblement déformées qui estime conjointement la séparation des sources et le décalage de phase entre source et référence, ainsi que son évaluation sur des mélanges artificiels.

Publications durant la thèse Les travaux sur les aspects de séparation guidée ont abouti à la publication d'un article dans une revue internationale à comité de lecture [161] et de deux articles de conférences internationales [160, 162]. Deux logiciels en lien avec cette thèse ont également été rendus publics à l'occasion de deux conférences internationales [34, 151]. Enfin, la collaboration avec l'équipe de recherche Linkmedia de l'Irisa a mené à la publication d'un article plus exploratoire sur un sujet connexe [77].

1.7 Plan du manuscrit et pistes de lectures

Ce travail se positionnant à l'intersection entre les deux champs de recherche que sont la séparation de sources et la détection de motifs, la **partie I** commence par détailler l'état de l'art de ces deux domaines dans les **chapitres 2 et 3**. La perspective d'une approche conjointe implique de nouvelles limitations qui sont soulignées dans les parties 2.3 et 3.4.4. Le **chapitre 4** décrit ensuite les approches de séparation de sources informée ainsi que d'autres approches de séparation exploitant la redondance, avant de les positionner par rapport à notre approche.

La **partie II** présente les différentes contributions scientifiques réparties dans trois chapitres distincts. Le **chapitre 5** porte sur la détection robuste de motifs et propose une nouvelle méthode de comparaison adaptée à la présence de plusieurs sources. Les **chapitres 6 et 7** regroupent les différentes méthodes de séparation guidée par signal de référence proposées. Le chapitre 6 s'attache à décrire un modèle général de déformation et à présenter différentes utilisations de ce modèle, alors que le chapitre 7 se focalise sur l'utilisation de références (typiquement de musique) qui sont déformées à une plus petite échelle, par exemple suite à un copier-coller et à l'ajout d'autres sources de manière analogique.

Enfin, la **partie III** est composée de deux chapitres. Le **chapitre 8** décrit la collaboration scientifique avec le Studio Maia qui est le projet dans lequel s'inscrit cette thèse. Sont notamment présentés les travaux de cette thèse ayant été transférés, ainsi que l'interface utilisateur et les autres outils développés pendant le projet. Une liste des autres applications potentielles de ces travaux est ensuite dressée, aussi bien pour les

professionnels que pour le grand public. Finalement, le **chapitre 9** conclut ce manuscrit et présente les perspectives scientifiques.

Première partie

État de l'art

Chapitre 2

Détection de motifs

Les techniques de détection de motifs audio sont des outils permettant de détecter une requête audio $x(t)$ dans un contenu audio $y(t)$. Elles sont en particulier adaptées au cas où la requête se répète de façon déformée dans le contenu, par exemple un mot prononcé par un autre locuteur, ou encore une reprise de la même musique. Ces traitements présentent un intérêt dans les situations où le contenu que l'on parcourt est d'une taille conséquente ou bien lorsqu'il faut comparer des scènes sonores complexes. La détection de motifs audio repose sur l'étape élémentaire de comparaison de deux segments audio pouvant être subdivisés en éléments plus petits, par exemple des trames de 100 millisecondes. Ce genre de comparaisons est aussi utile dans bien d'autres traitements audio plus complexes, comme par exemple en séparation de sources.

Dans ce chapitre, je commencerai par présenter cette étape de comparaison de segments audio avant d'introduire un modèle de déformation temporelle. Je parlerai ensuite des limites des techniques existantes avant de dresser un panorama des autres approches permettant de retrouver des requêtes ou des redondances dans un contenu audio.

2.1 Généralités

Vecteurs caractéristiques ou *feature vectors* Qu'il s'agisse de trames ou de signaux plus longs, la comparaison de deux segments audio peut se faire dans le domaine temporel. Cependant, il est préférable d'utiliser des représentations plus succinctes et adaptées aux éléments sonores que l'on cherche à comparer. En effet, les formes d'ondes des signaux à comparer peuvent être différentes, bien qu'elles semblent identiques pour une oreille humaine.

Ainsi, on préfère représenter les trames $x(t : t+w)$ et $y(t : t+w)$ de longueur w par leurs vecteurs caractéristiques

$$\mathsf{x}_n = \mathsf{x}_{n1},\ldots,\mathsf{x}_{nf},\ldots,\mathsf{x}_{nF} \in \mathbb{R}^F \tag{2.1}$$

$$\mathsf{y}_n = \mathsf{y}_{n1},\ldots,\mathsf{y}_{nf},\ldots,\mathsf{y}_{nF} \in \mathbb{R}^F \tag{2.2}$$

de taille F qui regroupent certaines propriétés de ces trames. Dans le cas de la détection de motifs, on s'intéresse aux deux séquences de vecteurs caractéristiques X et Y qui

représentent respectivement $x(t)$ et $y(t)$ dans leur intégralité :

$$\mathsf{X} = \mathsf{x}_1,...,\mathsf{x}_n,...,\mathsf{x}_{N_X} \in \mathbb{R}^{N_X \times F} \tag{2.3}$$

$$\mathsf{Y} = \mathsf{y}_1,...,\mathsf{y}_n,...,\mathsf{y}_{N_Y} \in \mathbb{R}^{N_Y \times F}, \tag{2.4}$$

où N_X et N_Y sont les nombres de trames. Les trames sont généralement superposées pour mieux rendre compte des propriétés du signal en chaque instant. Il est courant d'utiliser cinquante pour cent de superposition. Augmenter ce ratio mène généralement à une augmentation de la performance de détection mais aussi du temps de traitement.

Que ce soit pour la détection ou la séparation, j'utiliserai dans mes expériences différents types de *features* dont voici une brève description :

- Les ***Mel-Frequency Cesptral Coefficient* ou MFCC** sont calculés par transformée en cosinus discrète du spectre de log-puissance en échelle Mel. Ces *features* permettent de bien imiter la perception humaine de l'enveloppe spectrale. Ils sont couramment utilisés que ce soit pour la reconnaissance de la parole ou du locuteur ou la classification du genre musical [118] ou de sons environnementaux [114]. On utilise généralement treize coefficients ainsi que leurs dérivées premières et secondes par rapport au temps, soit 39 coefficients. Ils sont cependant peu robustes au bruit [100].
- Les douze **chroma features** permettent de représenter les douze demi-tons d'une octave musicale [53]. Ils sont obtenus par simple projection du spectre dans les zones fréquentielles correspondantes. Les notes de musique qui sont séparées d'une octave étant perçues comme semblables, cette représentation est bien adaptée pour représenter la perception humaine de la musique même si elle ne donne pas la fréquence fondamentale absolue. De façon plus générale, elle permet de bien résumer les trames de contenus musicaux.
- La ***Short Time Fourier Transform* (STFT)** (voir partie 3.1.2.1) permet de déterminer l'activité des fréquences sinusoïdales et leur phase au cours du temps.

On peut noter que la représentation STFT est couramment utilisée en séparation de sources et qu'elle est inversible, c'est-à-dire qu'il est possible de revenir dans le domaine temporel à partir de cette représentation contrairement aux MFCC et aux chroma.

Formulation usuelle La tâche de détection de motifs (*motif searching or spotting*) peut être exprimée comme « trouver tous les segments audio $\mathsf{Y}_{a:b} = \mathsf{Y}_a,...,\mathsf{Y}_b$ du contenu $y(t)$ qui sont suffisamment similaires à la représentation X de la requête $x(t)$ ». Ce qui peut être formulé comme

$$\mathsf{D}(\mathsf{X},\mathsf{Y}_{a:b}) < \mathsf{D}_{\text{seuil}} \tag{2.5}$$

avec D une distance globale, c'est-à-dire une mesure de dissimilarité dans le domaine *feature*-temporel. En ne considérant pas de déformation temporelle, c'est-à-dire par comparaison des trames deux à deux dans l'ordre chronologique, (2.5) se développe

comme

$$\sum_{n=1}^{N_X} \mathsf{d}(\mathsf{X}_n, \mathsf{Y}_{a+n-1}) < \mathsf{D}_{\text{seuil}} \tag{2.6}$$

avec nécessairement $N_X = b - a + 1$ et d une distance locale c'est-à-dire une mesure de dissimilarité dans le domaine des *features*.

Distance dans le domaine des *features* On définit la distance d entre deux vecteurs caractéristiques $\mathsf{x} \in \mathbb{R}^F$ et $\mathsf{y} \in \mathbb{R}^F$ comme une fonction de $\mathbb{R}^F \times \mathbb{R}^F$ dans $\mathbb{R}^+$. Plus x et y sont semblables, plus d est proche de 0.

Voici les principales distances existantes et que j'utiliserai dans la partie expérimentale :

— Les distances d_{l_p} issues des normes l_p du même nom :

$$\mathsf{d}_{l_p}(\mathsf{x},\mathsf{y}) = \left(\sum_{f=1}^{F} |\mathsf{x}_f - \mathsf{y}_f|^p \right)^{1/p}. \tag{2.7}$$

Cette famille de distances inclut les cas particuliers de la distance euclidienne ($p = 2$), de la distance de Manhattan ($p = 1$), du maximum de la différence ($p = +\infty$) et du nombre de différences non nulles ($p = 0$). Ainsi, plus p est grand moins le nombre de petites différences a d'importance et inversement plus p est petit moins les grandes différences ont d'importance.

— La « distance cosinus » [119] :

$$d_{\cos}(\mathsf{x},\mathsf{y}) = 1 - \frac{\mathsf{x}.\mathsf{y}}{\|\mathsf{x}\| \times \|\mathsf{y}\|}, \tag{2.8}$$

qui permet de rendre compte du fait que les vecteurs soient opposés (= 2), orthogonaux (= 1) ou colinéaires (= 0). C'est une distance normalisée, et lorsque l'on traite des vecteurs à valeurs positives elle prend des valeurs allant de 0 à 1.

— Les β-divergences qui sont définies plus loin par l'équation (3.21) [1] et qui sont généralement utilisées comme distances entre représentations temps-fréquence, par exemple pour mesurer la divergence entre deux STFT [20].

2.2 *Dynamic Time Warping* (DTW)

Nous avons vu précédemment un moyen simple (2.6) de comparer deux séquences de *features* sans inclure de déformation temporelle. Les techniques de *Dynamic Time Warping* (DTW) [97, 124] ont pour but d'apparier deux séquences par le biais d'un chemin de déformation. Ce chemin est typiquement obtenu par un algorithme de programmation dynamique [13] et notamment par l'utilisation d'une matrice de cumul des distances.

1. La distance euclidienne d_{l_2} fait aussi partie des β-divergences.

La DTW permet ainsi d'autoriser une déformation temporelle lors de la comparaison de deux séquences de *features* lors d'une tâche de recherche de motifs [118, 122]. La distance globale D est alors une distorsion cumulée (2.11). Cette technique d'alignement sera aussi utilisée dans le cadre de la séparation pour aligner le mélange et les références.

Afin de conserver des notations claires dans la description des étapes algorithmiques de la DTW, on note Z au lieu de $\mathsf{Y}_{a:b}$. Cela permet aussi d'avoir une notation plus générale et non spécifique à la détection de motifs.

Matrice de similarité La matrice de similarité est souvent utilisée pour représenter graphiquement le problème de recherche de chemin d'alignement. Il s'agit d'une matrice qui regroupe les comparaisons entre toutes les trames de deux séquences de *features* X et Z et qui s'écrit :

$$\mathsf{SM}_{(\mathsf{X},\mathsf{Z})}(n_x,n_z) = \mathsf{d}^{-1}(\mathsf{X}_{n_x},\mathsf{Z}_{n_z}) \forall n_x \in [1,N_X] \text{ et } n_z \in [1,N_Z], \tag{2.9}$$

où d^{-1} est l'inverse de la distance locale choisie. On peut par exemple y faire apparaitre un chemin de déformation (voir Figure 2.1c) ou dans le cas d'une matrice (symétrique) d'auto-similarité $\mathsf{SM}_{(\mathsf{X},\mathsf{X})}$ [119] voir apparaitre des redondances internes (voir Figure 2.1d). Ces deux figures ont été obtenues par comparaison de séquences de chroma représentant différentes occurrences du même leitmotiv musical d'un film. On peut voir sur la Figure 2.1c que le chemin optimal suit une « vallée » où les similarités sont fortes (en orange). De même, on voit apparaître sur la Figure 2.1d une diagonale principale et une sous-diagonale qui caractérise une répétition au sein du même exemple.

Des matrices de similarité seront aussi utilisées dans le cadre de la séparation pour pondérer des chemins d'alignement entre référence et mélange.

Chemin de déformation Un chemin de déformation est défini comme une succession de L paires de trames $p_k = (i_k,j_k)$:

$$P = \{p_1,...,p_k,...,p_L\} = \{(i_1,j_1),...,(i_L,j_L)\} = \{(i_k,j_k)\}_{k=1}^{L}, \tag{2.10}$$

qui suit :

- un critère de conditions aux limites : $p_1 = (1,1)$ et $p_L = (N_X,N_Z)$,
- et un critère de conditions de succession [2] : $\forall k \in [1,L-1], p_{k+1} = p_k + p_{\text{step}}$ avec $p_{\text{step}} \in P_{\text{step}} = \{(1,0),(0,1),(1,1)\}$.

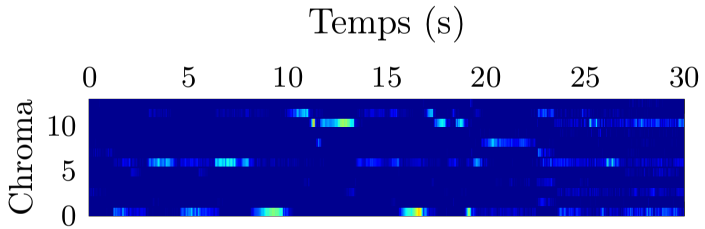
On peut noter que ces deux critères impliquent que $\forall k \in [1,L-1], i_k \leq i_{k+1}$ et $j_k \leq j_{k+1}$.

Le chemin P est alors évalué en terme de **distorsion cumulée** :

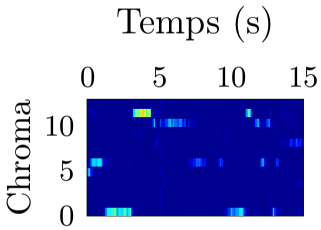
$$\mathsf{D}_P(\mathsf{X},\mathsf{Z}) = \sum_{k=1}^{L} \mathsf{d}(\mathsf{X}_{i_k},\mathsf{Z}_{j_k}). \tag{2.11}$$

Cette distorsion cumulée est relative à la distance locale d utilisée pour construire la matrice de similarité. Elle sert de fonction de coût à minimiser lorsque l'on cherche le

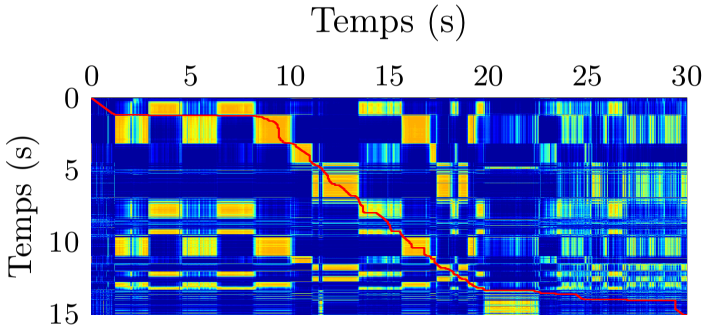
2. Il est aussi possible de définir d'autres ensembles de succession P_{step} [118, 120], mais je me restreindrai à la description et l'utilisation de celui-ci.



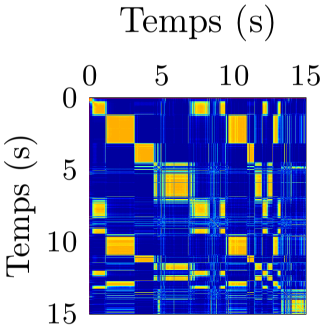
(a) Séquence de chroma Z



(b) Séquence de chroma X



(c) Exemple de matrice de similarité $\mathsf{SM}_{(\mathsf{X},\mathsf{Z})}$ et de chemin de déformation P (en rouge).



(d) Exemple de matrice d'auto-similarité $\mathsf{SM}_{(\mathsf{X},\mathsf{X})}$.

Figure 2.1 – Exemple de matrices de similarité. Le bleu désigne une faible intensité.

chemin optimal entre deux séquences X et Z :

$$\widehat{P} = \underset{P}{\operatorname{argmin}} \mathsf{D}_P(\mathsf{X},\mathsf{Z}). \tag{2.12}$$

C'est aussi cette distorsion cumulée qui sert de distance globale dans le cadre de la détection de motifs (2.5).

Matrice de cumul des distances Cet élément de programmation dynamique sera utilisé lors de la construction du chemin optimal. On définit la matrice de cumul des distances $ACU \in \mathbb{R}^{+N_X \times N_Z}$ relative aux séquences de *features* X et Z et à la distance de comparaison d comme

$$ACU(i,j) = \begin{cases} \sum_{k=1}^{j} \mathsf{d}(\mathsf{X}_1,\mathsf{Z}_{j_k}) & \text{si } i = 1 \\ \sum_{k=1}^{i} \mathsf{d}(\mathsf{X}_{i_k},\mathsf{Z}_1) & \text{si } j = 1 \\ \mathsf{d}(\mathsf{X}_i,\mathsf{Z}_j) + \min\limits_{p_{\text{step}} \in P_{\text{step}}} \{ACU((i,j) - p_{\text{step}})\} & \text{sinon.} \end{cases} \tag{2.13}$$

Cette matrice vérifie en particulier que $ACU(N_X,N_Z) = \min\{\mathsf{D}_P(\mathsf{X},\mathsf{Z})\}$ [118] et donne donc la garantie de trouver un chemin P optimal au sens de (2.12).

Construction du chemin optimal En fixant $p_L = (N_X,N_Z)$ à l'initialisation, on peut calculer le reste du chemin optimal à partir de ce couple en sens inverse de façon itérative. En d'autres termes, les p_{l-1} sont calculés à partir des couples précédents connus $p_l = (i,j)$ par la relation de récursivité suivante :

$$p_{l-1} = \begin{cases} (1,j-1) & \text{si } i = 1 \\ (i-1,1) & \text{si } j = 1 \\ p_l - \underset{p_{\text{step}} \in P_{\text{step}}}{\operatorname{argmin}} \{ACU(p_l - p_{\text{step}})\} & \text{sinon.} \end{cases} \tag{2.14}$$

2.3 Limitations des techniques de détection de motifs

Déformation Les déformations actuellement prises en compte par ces techniques de détection de motifs sont :

- les déformations auxquelles les *features* sont intrinsèquement invariantes, par exemple le changement de locuteur (resp. d'instrument ou d'octave) pour les MFCC (resp. les chromas).
- les déformations temporelles continues grâce à la DTW.

Les limitations sont donc la non-universalité des déformations pouvant être prises en compte et la contrainte de continuité du chemin d'alignement temporel imposée par la DTW. Ainsi, ce manque de flexibilité du chemin de déformation ne prévoit pas que l'élément à apparier soit entièrement masqué ou absent sur certaines trames.

Robustesse au bruit et présence d'autres sources La robustesse au bruit des techniques présentées se limite aux bruits stationnaires ou relativement faibles, alors que dans notre cas d'usage les bruits sont fortement non stationnaires (parole, musique) et potentiellement forts. Des *features* spécialement conçus dans ce but existent [100] mais restent limités tout comme les MFCC et les chromas.

Les distances couramment utilisées ne sont pas non plus adaptées à la parcimonie de ces bruits, c'est-à-dire au fait qu'ils présentent des pics d'intensité plutôt qu'un niveau constant. Par exemple, on utilise souvent les distances l_2 et l_1 dans le domaine des *features* (2.7) ou temporel (2.6) pour le calcul de la distorsion cumulée de la DTW (2.11).

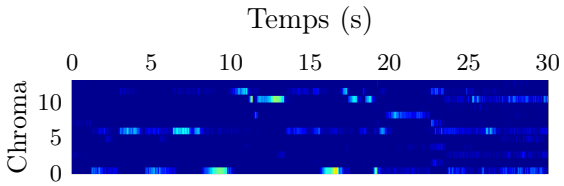
Exemple La Figure 2.2 illustre la non-robustesse des chroma-*features* et de la distance cosinus en présence de bruit. Comme pour la Figure 2.1c, on essaye d'apparier deux extraits musicaux provenant d'un film mais ici on ajoute un signal de parole à un des deux extraits musicaux. Ces figures montrent différents niveaux de bruits liés à cet ajout de parole. Le niveau est quantifié en terme de rapport signal-à-bruit (*Signal-to-Noise Ratio* ou SNR) en décibels (dB). Les lignes horizontales dans les figures sont dues à ces ajouts, le signal de parole ajouté contenant un certain nombre de silences. Les chemins noirs représentent la vérité terrain calculée à partir des deux exemples musicaux non bruités. C'est le même chemin que celui de la Figure 2.1c. Les chemins rouges sont les chemins estimés dans les conditions bruitées. La Figure 2.2 montre que plus l'ajout de parole est important plus l'estimation de la DTW est perturbée.

2.4 Approches ou tâches connexes

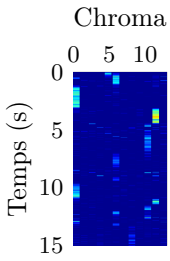
2.4.1 Détection de mots-clefs

De la même façon que pour la détection de motifs, les techniques de détection de mot-clef, aussi appelées *Keyword Spotting* ou KWS, recherchent des éléments audio dans un contenu à partir d'une requête. Cependant dans ce cas, la requête est sous forme symbolique (texte) et est représentée par un HMM. Les densités de probabilité des états du HMM sont généralement modélisées par des GMM appris au préalable sur un corpus de parole annotée. Les trames sont représentées par des vecteurs de descripteurs, par exemple des MFCC, et la requête est recherchée par décodage du HMM grâce à un algorithme de programmation dynamique [98, 164].

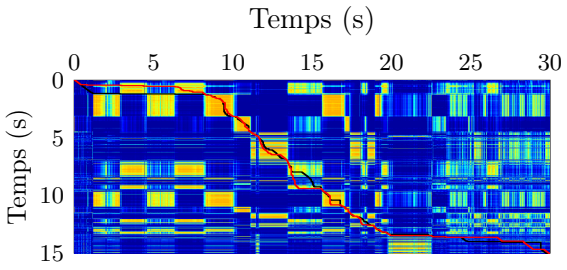
Limitations Par rapport aux techniques de détection de motifs, ces techniques sont plus robustes au bruit et à la distorsion de la parole, mais elles requièrent des ressources pour l'apprentissage, et sont encore moins rapides. Elles manquent aussi de flexibilité car elles sont restreintes au périmètre défini pendant la phase d'apprentissage (langue, lexique).



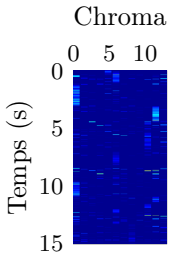
(a) Séquence de chroma Z (sans bruit)



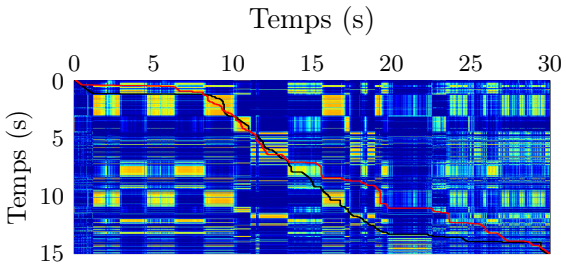
(b) Séquence de chroma X (SNR= 9 dB)



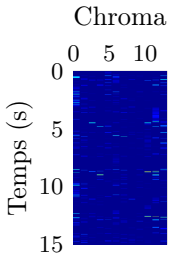
(c) Matrice de similarité $\mathsf{SM}_{(\mathsf{X},\mathsf{Z})}$ (SNR= 9 dB)



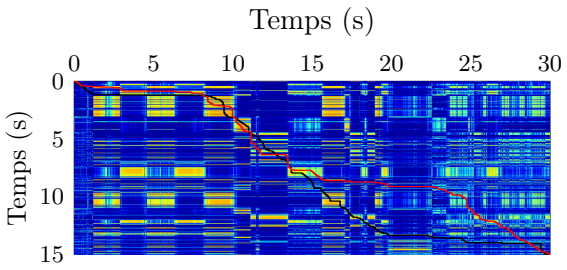
(d) Séquence de chroma X (SNR= 0 dB)



(e) Matrice de similarité $\mathsf{SM}_{(\mathsf{X},\mathsf{Z})}$ (SNR= 0 dB)



(f) Séquence de chroma X (SNR= −9 dB)



(g) Matrice de similarité $\mathsf{SM}_{(\mathsf{X},\mathsf{Z})}$ (SNR= −9 dB)

Figure 2.2 – Exemple de détérioration des performances de la DTW (en rouge) en présence d'une source de parole à différents niveaux de bruit. La vérité terrain est en noir.

2.4.2 Identification audio par empreinte acoustique

Les techniques d'identification audio par empreinte acoustique [38, 63, 64, 81, 179] aussi appelées *audio fingerprinting*, sont utilisées dans des applications d'identification de contenus audio identiques dans des grandes bases de recherche (morceaux de musique, enregistrements d'un même évènement). La requête est un signal sonore qui est répété dans le contenu de façon identique. Il subit cependant des déformations à l'échelle du signal dues aux canaux de transmission qu'il peut emprunter (radio, enceintes, milieu de propagation, réseau téléphonique, encodage) et peut comporter un bruit stationnaire de niveau modéré.

Le tâche d'identification est souvent basée sur une représentation succincte du signal (empreinte) et une table de hachage. La table de hachage permet de réduire le temps de recherche d'une empreinte dans une grande base de recherche [116]. Je ne détaillerai pas plus leur fonctionnement. En ce qui concerne les empreintes, l'utilisation de l'emplacement dans le spectrogramme de couples de pics d'intensité comme *features* [38, 64, 179] est l'approche la plus célèbre. Une autre approche [81] propose d'encoder chaque trame du signal par une courte série de bits. Chaque bit représente une bande large de fréquences et sa valeur dépend du signe de la variation temporelle de la différence avec la bande de fréquence supérieure.

Limitations Les techniques classiques d'*audio fingerprinting* sont robustes à l'égalisation ou la compression dynamique. Elles ont été récemment adaptées pour être robustes à de faibles variations de hauteur [64] ou à des changements de vitesse [46].

Bien qu'elle soient en théorie robustes à l'ajout de bruit, les performances s'effondrent en pratique pour des rapports signal-à-bruit inférieurs à -6 dB [179]. Ces techniques sont donc limitées lorsque il s'agit de musiques de film mixées en arrière-plan (environ -12 dB), et d'autant plus lorsque l'on cherche à apparier deux extraits avec ce niveau de bruit. De plus, un bruit non stationnaire comme de la parole, même faible, peut ajouter suffisamment de pics d'intensité pour corrompre l'empreinte.

Ces techniques ne sont pas non plus directement applicables à la parole car le signal correspondant à un mot diffère à chaque fois qu'il est prononcé.

2.4.3 Découverte de motifs

La découverte de motifs est la version non supervisée de la détection de motifs, c'est-à-dire sans requête X en entrée [88]. On recherche alors tous les motifs qui se répètent dans le contenu Y. Cette tâche est formulée par Muscariello [120] comme « trouver tous les couples de segments $\{a : b\}$, $\{c : d\}$ » tels que :

$$\mathsf{D}(\mathsf{Y}_{a:b}, \mathsf{Y}_{c:d}) < \mathsf{D}_{\text{seuil}} \tag{2.15}$$

$$|b - a| > L_{\min} \tag{2.16}$$

$$a < b < c < d \tag{2.17}$$

Après regroupement (*clustering*) des couples comportant un segment commun (ou proche), on obtient une librairie de motifs. Chaque motif présente donc au moins deux

occurrences. Selon les choix de distance, de *features* et de taille minimale du motif, un motif peut représenter un mot, un morceau de musique ou tout autre élément répété. La matrice d'auto-similarité de Υ est une des représentations graphiques qui permet le mieux de visualiser l'ensemble des motifs découverts.

Une première méthode de résolution consiste à comparer entre elles tous les segments possibles. C'est cependant une stratégie inadaptée à la plupart des situations. Selon la taille du contenu et le type de motifs recherchés, d'autres stratégies permettent de réduire le temps de traitement. Par exemple, l'utilisation d'une DTW avec des contraintes de condition aux limites relâchées (*segmental locally normalized DTW* [121, 122]) permet d'effectuer des comparaisons de segments plus petits (amorces) et d'étendre ces segments tant que le critère de comparaison est respecté. Une stratégie compatible est de réduire le voisinage de recherche pour ces amorces. Ainsi, avant qu'un motif soit placé dans la libraire courante, il doit présenter deux occurrences voisines [88]. On réduit ainsi le nombre de comparaisons.

Limitations Les techniques de comparaison élémentaire étant les mêmes que pour la détection de motifs, les mêmes limitations sont valables pour la découverte de motifs (distances, *features*) (voir partie 2.3). On peut cependant ajouter une limitation mineure de la formulation actuelle face aux signaux constitués de plusieurs sources. En effet, elle ne prévoit pas que deux motifs distincts apparaissent en même temps [88, 120].

Chapitre 3

Séparation de sources

En audio, la séparation de sources est la tâche qui consiste à retrouver les différents signaux (sources) qui composent un mélange observé. Le mélange et les sources peuvent présenter des caractéristiques très diverses qu'il convient de prendre en compte à la fois dans la formalisation du problème mais aussi dans la modélisation et l'estimation des grandeurs physiques de chaque source.

Dans ce chapitre, nous présenterons tout d'abord différentes formalisations du problème adaptées à différentes hypothèses et situations. Un tour d'horizon des approches existantes sera ensuite effectué avant de détailler l'approche qui sera principalement utilisée dans le reste de la thèse : la factorisation en matrices positives (*Nonnegative Matrix Factorization* ou NMF). Enfin, un état de l'art des techniques de séparation guidée et plus particulièrement de séparation guidée par signal de référence sera établi.

3.1 Problème

3.1.1 Formulation dans le domaine temporel

L'observation est un mélange audio multicanal $\mathbf{x}(t) = [x_1(t),...,x_I(t)]$ contenant I canaux indexés par i. On suppose que ce mélange est la somme des images spatiales $\mathbf{y}_j(t) = [y_{1j}(t),...,y_{Ij}(t)]$ de plusieurs sources indexées par $j \in \mathcal{J}$ [33] :

$$\mathbf{x}(t) = \sum_{j \in \mathcal{J}} \mathbf{y}_j(t) \text{ avec } \mathbf{x}(t),\mathbf{y}_j(t) \in \mathbb{R}^I. \tag{3.1}$$

L'image spatiale d'une source est l'ensemble des contributions de cette source au niveau des canaux du mélange. Dans le cas d'une source ponctuelle, c'est-à-dire d'un seul emplacement d'émission, la propagation entre cette source émettrice $s_j(t)$ et son image spatiale $\mathbf{y}_j(t)$ peut être formulée comme suit [115] :

$$\mathbf{y}_j(t) = \sum_{\tau} \mathbf{h}_j(\tau) s_j(t-\tau) \tag{3.2}$$

où $\mathbf{h}_j(\tau) = [h_{1j}(\tau),...,h_{Ij}(\tau)]^T \in \mathbb{R}^I$ est l'ensemble des filtres de propagation entre l'emplacement d'émission et les emplacements de réception.

3.1.1.1 Mélanges convolutifs, anéchoïques ou instantanés

Les filtres de propagation $h_{ij}(\tau)$ sont fréquemment modélisés par des filtres à réponse impulsionnelle finie (RIF). Dans le cas réverbérant, c'est-à-dire de plusieurs chemins de propagation, l'ordre du filtre (RIF) dépendra directement du temps de réverbération qui est une caractéristique propre à l'environnement de propagation. On parle alors de modèle convolutif en référence à l'équation (3.2).

En environnement anéchoïque, seul le chemin direct de propagation est possible. Cette situation peut être modélisée par un retard τ_{ij} et une atténuation a_{ij} :

$$h_{ij}(\tau) = a_{ij}\delta(\tau - \tau_{ij}). \tag{3.3}$$

Enfin, le cas linéaire instantané est le cas où le filtre représente une simple atténuation :

$$h_{ij} = a_{ij}. \tag{3.4}$$

Ces deux derniers cas ne représentent pas une configuration physique réelle de propagation, mais une forme synthétique de mixage.

3.1.1.2 Sources diffuses

On peut également considérer une source comme diffuse. Cela signifie qu'elle n'émet pas d'un unique point mais d'une zone. La source peut alors être assimilé à un ensemble $\mathcal{Q}_j$ de sous-sources ponctuelles. L'image d'une telle source j est alors la somme des contributions provenant des différentes sous-sources et l'équation (3.2) devient :

$$y_{ij}(t) = \sum_{q \in \mathcal{Q}_j} (h_{iq} * s_q)(t). \tag{3.5}$$

Cela peut par exemple représenter le cas des caisses de résonance de certains instruments de musique.

3.1.2 Formulation dans le domaine temps-fréquence

Les algorithmes de séparation de sources audio opèrent généralement dans le domaine temps-fréquence et leur but est d'estimer la contribution de chaque source en chaque point temps-fréquence dans le but de l'extraire du mélange.

3.1.2.1 Représentation temps-fréquence

Une représentation temps-fréquence est calculée pour chaque canal i à partir du signal temporel correspondant $x_i(t)$ et donne une estimation de l'activité de certaines fréquences f à certains instants n. On utilise généralement la Transformée de Fourier à Court Terme (*Short Time Fourier Transform* ou STFT) [3] dont le calcul peut se faire par application d'un banc de filtres passe-bande et fenêtrage temporel. Cependant, on utilise en pratique la transformée de Fourier rapide (*Fast Fourier Transform* ou

FFT). Les fenêtres temporelles sont généralement superposées pour ne pas induire de discontinuité. On obtient un ensemble de coefficients temps-fréquence $x_{fn} \in \mathbb{C}$.

Il existe des variantes de cette représentation où l'espacement entre les fréquences suit une échelle autre que linéaire. La transformée à Q constant (*Contant Q Transform* ou CQT) suit une échelle logarithmique adaptée aux harmoniques musicales et est exploitée par des modèles invariants par translation [68, 87, 126, 155]. On peut aussi utiliser l'échelle ERB [149] adaptée à la perception auditive humaine.

3.1.2.2 Modèle gaussien local

Dans le domaine STFT, l'équation (3.1), peut s'écrire :

$$\mathbf{x}_{fn} = \sum_{j \in \mathcal{J}} \mathbf{y}_{j,fn} \tag{3.6}$$

où $\mathbf{x}_{fn} = [x_{1,fn},...,x_{I,fn}] \in \mathbb{C}^I$, $\mathbf{y}_{j,fn} = [y_{1j,fn},...,y_{Ij,fn}] \in \mathbb{C}^I$, et $f = 1,...,F$ et $n = 1,...,N$ sont respectivement les indices fréquentiels et temporels de la STFT. Une hypothèse usuelle (aussi appelé modèle gaussien local) est alors de considérer que les coefficients STFT des images spatiales des sources $\mathbf{y}_{j,fn}$ suivent une distribution gaussienne centrée [136, 175] :

$$\mathbf{y}_{j,fn} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{y}_{j,fn}}) \tag{3.7}$$

dont la covariance se factorise en un terme scalaire de puissance spectrale $v_{j,fn} \in \mathbb{R}_+$ et une matrice de covariance spatiale $\mathbf{R}_{j,fn} \in \mathbb{C}^{I \times I}$:

$$\mathbf{\Sigma}_{\mathbf{y}_{j,fn}} = v_{j,fn} \mathbf{R}_{j,fn}. \tag{3.8}$$

Cette dernière modélise les caractéristiques spatiales des sources comme les différences de phase ou d'intensité et la corrélation entre les différents canaux.

Classiquement, l'hypothèse de bande étroite [43] suppose que la matrice de covariance spatiale est invariante en temps, de rang 1 et se factorise comme :

$$\mathbf{R}_{j,f} = \mathbf{h}_j(f) \mathbf{h}_j(f)^H, \tag{3.9}$$

où $\mathbf{h}_j(f)$ est la transformée de Fourier du filtre de mélange $\mathbf{h}_j(\tau)$. C'est une hypothèse raisonnable lorsque les sources ou leur environnement varient peu au cours du temps, que le mélange est instantané ou que les sources sont ponctuelles et faiblement réverbérées. On comprend donc qu'elle comporte certaines limitations dans des situations plus réelles.

Cette hypothèse historique a été remise en cause par les travaux de Duong [43, 45] qui propose l'utilisation de matrices de covariance spatiale de rang plein pour répondre au cas réverbéré ou diffus. Une généralisation de cette idée peut être écrite comme [136] :

$$\mathbf{R}_{j,f} = \mathbf{A}_{j,f} \mathbf{A}_{j,f}^H, \tag{3.10}$$

et découle de (3.5). $\mathbf{A}_{j,f} \in \mathbb{C}^{I \times R_j}$ et R_j est le rang des matrices de mélanges $\mathbf{R}_{j,f}$ et $\mathbf{A}_{j,f}$. Les matrices sont de rang plein lorsque $R_j = I$.

3.1.2.3 Autres indices spatiaux

Certains travaux utilisent d'autres représentations des caractéristiques spatiales encodés par la matrice de covariance spatiale. On peut par exemple observer la **différence intercanale de phase**

$$arg\Big(\frac{x_{i,fn}}{x_{i',fn}}\Big) \tag{3.11}$$

et la **différence intercanale d'amplitude** :

$$\frac{|x_{i,fn}|}{|x_{i',fn}|}. \tag{3.12}$$

On peut aussi citer la **cohérence intercanale** [9] qui est définie entre deux canaux i et i' en chaque point temps-fréquence.

3.1.2.4 Filtrage temps-fréquence

L'estimée de l'image spatiale d'une source en un point temps-fréquence $\hat{\mathbf{y}}_{j,fn}$ est généralement obtenue par application d'un filtre multicanal $\mathbf{G}_{j,fn} \in \mathbb{C}^{I\times I}$ au mélange $\mathbf{x}_{fn}$:

$$\hat{\mathbf{y}}_{j,fn} = \mathbf{G}_{j,fn}\mathbf{x}_{fn}. \tag{3.13}$$

Les signaux temporels correspondants à ces images peuvent ensuite être reconstruits par transformée de Fourier inverse ou *Inverse Short Time Fourier Transform* ou ISTFT.

Masque binaire Une première hypothèse est qu'une seule source est présente en chaque point temps-fréquence [128, 184] :

$$\forall(i,i'),\ \mathbf{G}_{j,fn}(i,i') = \begin{cases} 1 & \text{si } i = i' \text{ et que la source } j \text{ domine} \\ 0 & \text{sinon.} \end{cases} \tag{3.14}$$

Ce genre de filtre appelé masque binaire est simple d'utilisation et peut être construit par classification. Cependant il introduit des erreurs de discontinuité ou l'apparition d'harmoniques fantômes.

Filtre de Wiener Une autre possibilité qui tend à devenir l'approche de référence est l'utilisation d'un filtre de Wiener généralisé [181] qui permet l'estimation des sources de façon optimale au sens des moindres carrés (*Minimum Mean Square Error* ou MMSE). Dans le cas gaussien (3.7), le filtre à appliquer est donnée par l'équation :

$$\mathbf{G}_{j,fn} = \mathbf{\Sigma}_{\mathbf{y}_{j,fn}}\mathbf{\Sigma}^{-1}_{\mathbf{x}_{fn}} \tag{3.15}$$

où $\mathbf{\Sigma}_{\mathbf{x}_{fn}} = \sum_{j\in\mathcal{J}} \mathbf{\Sigma}_{\mathbf{y}_{j,fn}}$. On peut trouver dans la littérature d'autres formulations de ce filtre adaptées aux signaux localement stationnaires [14] ou aux représentations temps-fréquence quadratiques [136].

3.1.3 Évaluation

L'évaluation de la séparation se fait par comparaison à la vérité terrain des signaux reconstruits par ISTFT à partir des images.

3.1.3.1 Métriques pour l'évaluation de la séparation

Les métriques objectives couramment utilisées sont le rapport Signal-à-Distorsion (*Signal-to-Distortion Ratio* ou SDR), le rapport Signal-à-Artéfacts (*Signal-to-Artifacts Ratio* ou SAR) et enfin le rapport Signal-à-Interférence (*Signal-to-Interference Ratio* ou SIR), tous trois disponibles dans la *toolbox* Matlab BSS Eval [174]. Ce sont ces métriques qui seront utilisées pour l'évaluation dans le domaine temporel des algorithmes et approches proposés dans cette thèse.

3.1.3.2 Métriques pour l'évaluation des modèles

La divergence d'Itakura-Saito (IS) [65] est une mesure de comparaison de spectrogrammes généralement utilisée comme critère de convergence dans l'estimation des modèles de spectrogramme. Elle sera également utilisée comme métrique d'évaluation dans certaines situations particulières.

3.1.3.3 Performances oracle

Le terme « oracle » fait référence à plusieurs situations où les vraies sources sont utilisées durant la séparation. Cela permet d'estimer des bornes de performances pour les différentes étapes (estimation du modèle, filtrage) d'une approche donnée.

On peut par exemple parler de **filtrage oracle**. Il s'agit alors de la performance de séparation obtenue à partir du filtre de séparation construit en observant les vraies sources. Le masque binaire oracle peut être construit par comparaison de l'intensité des vraies sources en chaque point temps-fréquence. Les erreurs déjà précitées apparaissent également dans ce cas et sont inhérentes à l'utilisation de ce type de filtre. Dans le cas du filtre de Wiener qui est optimal, le filtre oracle mène à une séparation presque parfaite. L'évaluation d'un filtre oracle permet donc de quantifier l'erreur introduite par un type de filtre donné et de placer une première borne supérieure de performance.

L'utilisation d'un modèle pour chaque source introduit également une erreur d'estimation. On parle alors d'**estimation oracle d'un modèle**. Il s'agit d'estimer les paramètres optimaux du modèle qui approximent la vraie source, puis de reconstruire le signal temporel. La performance oracle obtenue donne une borne supérieure de performance du modèle et est nécessairement inférieure à celle de l'oracle de filtrage.

3.2 Les approches

La partie précédente a présenté différentes formulations du problème de séparation. Nous allons maintenant nous intéresser aux approches permettant de résoudre ce problème.

Une approche de séparation comprend généralement un modèle pour les sources, une méthode d'estimation des paramètres de ce modèle et un type de filtrage. Le choix de l'approche est intimement lié à la formulation du problème, à l'indétermination de celui-ci (c'est-à-dire au nombre de sources par rapport au nombre de canaux), à la présence de données extérieures et aux différentes hypothèses faites sur les sources.

Je donne ci-après de façon non exhaustive quelques différences notables entre approches ainsi qu'une brève description des approches les plus connues.

3.2.1 Quelques dichotomies

Sur/sous-déterminé Historiquement, le problème de la séparation sur-déterminée (c'est-à-dire avec plus de canaux que de sources) a commencé à être exploré au début des années 90 par différentes communautés. Une communauté transverse s'est alors structurée jusqu'à l'apparition des premières conférences au début des années 2000. Avec l'arrivée à maturation, cette communauté a progressivement attribué de plus en plus d'importance au cas sous-déterminé et les premières campagnes d'évaluation en audio ont commencé à la fin de la même décennie [176].

L'équivalence entre l'estimation de la matrice de mélange et l'estimation des sources est la propriété majeure du cas sur-déterminé. Au contraire, un problème de séparation sous-déterminée nécessitera l'estimation des sources et de la matrice de mélange. Ces deux approches sont donc foncièrement différentes. En particulier, les approches sur-déterminées ne sont pas applicables dans le cas mono-canal.

Aveugle/guidée/informée Les approches dites aveugles n'ont recours à aucun modèle avancé ni à des données extérieures permettant un apprentissage préalable. Lorsque l'on incorpore de l'information extérieure, on parle d'approche guidée (voir partie 3.4 pour plus de détails) et, lorsque cette information est obtenue à partir des vraies sources, on parle d'approche informée (voir partie 4.1 pour plus de détails).

On trouve aussi dans la littérature le terme « semi-aveugle » qui désigne une approche très faiblement guidée, ainsi que le terme « semi-informé » qui désigne le cas guidé [110]. Le terme « informée » a aussi été parfois utilisé dans le passé [111] pour parler d'approche guidée.

Déterministe/probabiliste Les approches déterministes cherchent à approcher un objectif quantitatif. En séparation de sources audio, cela revient par exemple à minimiser une divergence entre deux représentations temps-fréquence. On utilise typiquement pour cela des techniques d'optimisation pouvant inclure des contraintes.

À l'opposé, les approches probabilistes sont des méthodes d'inférence statistique dans lesquelles les sources sont modélisées par des variables aléatoires. On peut alors injecter des connaissances à propos des sources sous la forme de lois de probabilité à priori. On cherche ensuite à maximiser la vraisemblance des paramètres de ces lois à partir du mélange observé. Dans le cas de la séparation de sources, on parle plus souvent de modèle bayésien, les lois de probabilités étant exprimées conditionnellement au

mélange ou aux autres variables aléatoires. On peux aussi parler d'approches probabilistes aveugles lorsque les lois à priori sont définies sans avoir recours à des données extérieures.

Hypothèse gaussienne/parcimonieuse Ces deux hypothèses statistiques sont certainement les plus courantes et concernent les aspects spectraux des sources. Elles sont parfois mises en opposition sans pour autant être complémentaires, l'hypothèse de parcimonie accompagnant souvent une hypothèse d'indépendance statistique des sources et de non-gaussianité.

Faire une hypothèse de parcimonie [78, 142] revient à considérer que seulement un faible nombre de zones temps-fréquence ont une énergie significative. On peut justifier cette hypothèse par le fait que les signaux audio concentrent leur énergie dans des structures parcimonieuses comme des harmoniques ou encore des impulsions rythmiques. Cela peut par exemple se traduire par la modélisation des sources par des distributions à queues lourdes ou bien l'utilisation d'un masquage binaire (3.14).

L'hypothèse gaussienne précitée est uniquement exploitée par des approches probabilistes alors que l'hypothèse de parcimonie est aussi exploitable par des approches déterministes, typiquement des techniques d'optimisation cherchant à minimiser la norme l_0 ou l_1.

Information spatiale/spectrale Nous avons vu dans la partie précédente qu'il existe un certain nombre d'informations spatiales exploitables dans le cas multicanal. Si on fait une hypothèse de parcimonie, ces informations sont suffisantes pour regrouper les points temps-fréquence par des techniques simples de *clustering*.

En revanche, concernant l'information spectrale, on ne dispose que de l'énergie en chaque point temps-fréquence, ce qui n'est pas une donnée discriminante des sources. La discrimination des sources est dans ce cas réalisable par observation et structuration de l'ensemble des points temps-fréquence par des techniques plus avancées ou qui incluent plus d'informations à propos du spectre des sources. Je donnerai plus de détails sur certaines de ces approches dans les deux parties suivantes.

Enfin, on peut noter que certaines approches se focalisent uniquement sur l'un des deux types d'informations alors que d'autres cherchent à les exploiter conjointement [136].

3.2.2 Exemples d'approches spatiales

Je présente ci-après une série d'approches que l'on peut qualifier de « spatiales » de par leur filiation commune aux techniques de formation de voies (*beamforming*). La liste suivante est ordonnée par ordre chronologique d'apparition des approches.

Beamforming Le *beamforming* [22, 168] sépare les sources selon leur direction d'arrivée par filtrage spatial. Ces techniques se sont appuyées sur les théories déjà existantes pour les réseaux d'antennes, que ce soit pour l'approximation de filtres inverses, ou pour le filtrage selon la direction d'arrivé.

ICA L'analyse en composantes indépendantes (*Independent Component Analysis* ou ICA) est un groupe d'approches qui estiment le filtre inverse en chaque point temps-fréquence en supposant que les sources sont statistiquement indépendantes les unes des autres (non-gaussienne et i.i.d.) ou qu'elles sont gaussiennes et non stationnaires. Plus concrètement on cherche à minimiser une approximation de l'information mutuelle entre les signaux de sortie. On peut noter que cela est équivalent à une estimation au maximum de vraisemblance si on fixe la distribution à priori des sources [32].

Ce genre d'approches reste limité au cas linéaire et sur-déterminé, l'ensemble des filtres de propagation formant un système non inversible dans le cas sous-déterminé [18].

SCA L'analyse en composantes parcimonieuses (*Sparse Component Analysis* ou SCA) est aussi un groupe d'approches qui font les mêmes hypothèses que l'ICA (sauf l'indépendance) en ajoutant une hypothèse de parcimonie dans le plan temps-fréquence ce qui permet de s'attaquer au cas sous-déterminé. Cela a donné naissance entre autres à DUET et DEMIX.

DUET La technique connue sous le nom de *Degenerate Unmixing Estimation Technique* (DUET) [96, 184] est une technique de SCA qui regroupe les points temps-fréquence sur la base de leurs différences de temps d'arrivée (équivalent à (3.11)) et d'intensité entre canaux (3.12). Un masquage binaire est ensuite appliqué. DUET n'a pas été conçue pour traiter des mélanges convolutifs.

DEMIX La technique appelée *Direction Estimation of Mixing matrIX* (DEMIX) [4,5] est aussi une technique de SCA et ajoute aux hypothèses de DUET que les sources sont présentes seules dans des zones temps-fréquence élargies. Cette approche utilise un algorithme de *clustering* séquentiel pour suivre cette hypothèse ainsi que des mesures de confiance de l'estimation des matrices de mélange.

Full-Rank model Les travaux de Duong [43, 45] sur les matrices de covariance $\mathbf{R}_{j,f}$ de rang plein (3.10) permettent une modélisation plus rigoureuse des mélanges convolutifs et des sources diffuses. Je présenterai dans la partie 3.3.2.2 un algorithme EM de l'état de l'art capable d'estimer de telles matrices de rang plein.

3.2.3 Approches spectrales célèbres

Contrairement aux approches spatiales, les approches spectrales ne se basent pas sur l'information intercanale. Elles sont par nature capables de traiter des mélanges mono-canal ou sous-déterminés.

NMF La factorisation en matrices positives (*Nonnegative Matrix Factorization* ou NMF) [20, 21] permet de modéliser un spectrogramme comme le produit d'un dictionnaire de composantes fréquentielles et d'une matrice d'activations temporelles. Il est largement utilisé en séparation de sources pour la modélisation de chaque source. La

partie suivante est entièrement consacrée aux approches de l'état de l'art qui découlent de ce modèle.

Spectral GMM Les modèles de mélange de gaussiennes dans le domaine spectral [14, 15,134] sont une alternative antérieure à la formulation locale du modèle gaussien (3.7). Les trames $\mathbf{y}_{j,n}$ de chaque source j sont des vecteurs colonne de F éléments qui suivent un mélange de K gaussiennes, ce qui en mono-canal peut être formulé comme

$$\mathbf{y}_{j,n} \sim \sum_k \pi_{jk} \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Sigma}_{jk}) \tag{3.16}$$

où $\sum_k \pi_{jk} = 1$ et $\mathbf{\Sigma}_{jk} = \mathbf{diag}([v_{j,kf}]_f)$. Cette formulation essaye de rendre compte de la structure fréquentielle globale des sources. Il existe une version normalisée [6] qui permet de régler les problèmes d'amplitude par l'ajout aux variances des gaussiennes d'un terme dépendant du temps.

PLCA L'analyse probabiliste en composantes latentes (*Probabilistic Latent Component Analysis* ou PLCA) [12,72,153] considère l'ensemble des points temps-fréquence comme un histogramme issu du tirage d'une loi jointe de variables aléatoires de fréquence (f) et de temps (n). Ces variables sont supposées indépendantes conditionnellement à une variable latente (k), ce qui conduit à l'écriture suivante de la loi de probabilité jointe :

$$v_{j,fn} \sim P(f,n) = \sum_k P(k)P(f|k)P(n|k). \tag{3.17}$$

L'estimation de ce modèle se fait la plupart du temps au sens du maximum de vraisemblance et la présence de variables cachées rend l'utilisation de l'algorithme EM particulièrement naturelle. Lorsque la variable latente représente une composante de dictionnaire, l'estimation du maximum de vraisemblance de ce modèle est équivalente à une NMF avec comme critère de coût la divergence de Kullback-Leibler (KL). Il existe aussi une variante de cette approche qui est invariante par translation [158].

REPET La *REpeatting Pattern Extraction Technique* (REPET) [144, 146] et ses extensions [112, 145] sont des techniques de modélisation de spectrogrammes basées sur la répétition de l'accompagnement musical. Elles seront décrites et comparées à l'approche SPORES dans les parties 4.2 et 4.3.

3.3 Factorisation en matrices positives

3.3.1 Généralités

En audio, la factorisation en matrices positives (NMF) [65, 107, 156] permet de modéliser le spectrogramme de puissance $X = [\|x_{fn}\|^2]_{f,n} \in \mathbb{R}_+^{F \times N}$ d'un signal $x(t)$ par le

produit d'un dictionnaire de composantes fréquentielles $W \in \mathbb{R}_+^{F\times K}$ et d'une matrice d'activations temporelles $H \in \mathbb{R}_+^{K\times N}$:

$$X \approx V = WH, \tag{3.18}$$

ou encore :

$$x_{fn} \approx v_{fn} = \sum_{k=1}^{K} w_{fk} h_{kn} \tag{3.19}$$

où k sont les indices des composantes du dictionnaire. L'utilisation de cette technique de réduction de dimension ($K\times(F+N) << F\times N$) permet une représentation du spectrogramme en éléments simples et s'est généralisée en audio au cours des dix dernières années, notamment en séparation de sources [20, 47, 72, 84, 125].

Les β-divergences comme fonction de coût Une façon déterministe d'approximer X par V est de chercher à minimiser une divergence D entre X et V :

$$D(X|V) = D(X|WH) = \sum_{f,n} d([X]_{fn}|[WH]_{fn}), \tag{3.20}$$

où $d(x|y) \geq 0$ et $d(x|y) = 0$ si et seulement si $x = y$.

En ce qui concerne la factorisation positive des spectrogrammes audio, on utilise généralement la classe des β-divergences définie comme [20, 65, 84, 86] :

$$d_\beta(x|y) \stackrel{\text{def}}{=} \begin{cases} \frac{x^\beta+(\beta-1)y^\beta-\beta x y^{\beta-1}}{\beta(\beta-1)} & \text{pour} \quad \beta \in \mathbb{R}\setminus\{0,1\} \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \text{pour} \quad \beta = 0 \\ x\log(\frac{x}{y}) + (y-x) & \text{pour} \quad \beta = 1 \end{cases} \tag{3.21}$$

et qui regroupe les divergences d'Itakura-Saito (IS) ($\beta = 0$), de Kullback-Leibler (KL) ($\beta = 1$) et euclidienne ($\beta = 2$). Toutes les $d_\beta(x|y)$ présentent un unique minimum pour $y = x$, ce qui justifie leur utilisation comme fonction de coût.

Équivalence probabiliste Il a été démontré [65] que la minimisation déterministe de la divergence d'IS $D_{IS}(V|WH)$ équivaut à l'estimation au sens du maximum de vraisemblance des paramètres de NMF dans le cas où le signal suit le modèle gaussien complexe (3.7) et sa variance est modélisée par (3.19). De même, la divergence de KL correspond [153] quant à elle au cadre probabiliste posé par la PLCA (3.17) et la divergence euclidienne aux modèles de bruit gaussien additif [152] à valeurs réelles que je ne présente pas dans ce manuscrit.

NMF et séparation En séparation de sources, la NMF est souvent utilisée pour modéliser le spectrogramme des sources comme $V_j = W_j H_j$. Dans le cas mono-canal (voir partie 3.3.2.2 pour le cas multicanal), la fonction de coût à minimiser est alors :

$$D\Big(X\Big|\sum_{j\in\mathcal{J}} V_j\Big) = D\Big(X\Big|\sum_{j\in\mathcal{J}} W_j H_j\Big), \tag{3.22}$$

où $X = [\|x_{fn}\|^2]_{fn}$ est le spectrogramme de puissance du mélange mono-canal.

3.3.2 Les algorithmes réutilisés dans ce manuscrit

Il existe bon nombre d'algorithmes capable d'estimer les paramètres de NMF [36]. On peut en trouver un panorama dans les manuscrits de thèse de Hennequin [84] et Bertin [20]. Je me restreins ici à la description de deux algorithmes usuels qui se basent sur l'hypothèse gaussienne (3.7) et qui seront réutilisés plus tard dans le manuscrit. En particulier, nous adapterons ces algorithmes aux modèles NMF source-filtre (voir partie 3.3.3) et NMF conjointe (voir chapitre 6).

3.3.2.1 Mises à jour multiplicatives simples

En posant comme modèle spectral $V = WH$, le gradient de la β-divergence selon le paramètre W est [65]

$$\nabla_W D_\beta(X|WH) = ((WH)^{\cdot[\beta-2]} \odot (WH - X))H^T \tag{3.23}$$

où $\odot$ représente l'opérateur de multiplication point à point et $V^{\cdot[-p]}$ représente la matrice ayant pour éléments $(V_{ij})^{-p}$. Après avoir écrit ce gradient comme la somme d'un terme négatif $[\nabla_W]_- = -((WH)^{\cdot[\beta-2]} \odot X)H^T$ et d'un terme positif $[\nabla_W]_+ = (WH)^{\cdot[\beta-1]}H^T$, le paramètre W peut alors être mis à jour par multiplication par le rapport de ce terme négatif et de ce terme positif, ce qui s'écrit comme

$$W \leftarrow W \odot \frac{[\nabla_W]_-}{[\nabla_W]_+} = W \odot \frac{((WH)^{\cdot[\beta-2]} \odot X)H^T}{(WH)^{\cdot[\beta-1]}H^T} \tag{3.24}$$

et de la même façon on obtient la mise à jour pour le paramètre H

$$H \leftarrow H \odot \frac{W^T((WH)^{\cdot[\beta-2]} \odot X)}{W^T(WH)^{\cdot[\beta-1]}}. \tag{3.25}$$

Cette mise à jour garantit la diminution de la β-divergence [66]. Ainsi, en itérant alternativement (3.24) et son équivalent pour H, V se rapprochera de X et convergera vers un optimum local pour chaque source j.

Dans le cas de la séparation de sources, c'est-à-dire avec (3.22) comme fonction de coût, on obtient les mêmes mises à jour.

3.3.2.2 GEM multicanal

Différentes approches ont adapté la NMF pour qu'elle puisse traiter des mélanges multicanaux dans le cas convolutif [48, 66, 130] et par la suite dans le cas diffus [45, 136]. Je rappelle ci-après l'algorithme d'espérance-maximisation généralisé (*Generalized Expectation-Maximization* ou GEM) proposé dans [136] qui fait l'hypothèse d'un modèle gaussien complexe (3.7) multicanal (3.8) et de matrice de covariance de rang plein (3.10).

Les algorithmes *Expectation-Maximization* ou EM [40] sont particulièrement adaptés pour les problèmes d'estimation au sens du maximum de vraisemblance où cohabitent des données observées $\mathbf{X}$ et non observées $\mathbf{S}$. Ces algorithmes itératifs alternent une étape d'estimation (E-step) qui calcule l'espérance ($\mathbb{E}$) de la log-vraisemblance

$\mathbb{E}_{\mathbf{Z}|\theta^c}[\log p(\mathbf{Z}|\theta)] \stackrel{\Delta}{=} Q(\theta,\theta^c)$ des données complètes $\mathbf{Z} = \{\mathbf{X},\mathbf{S}\}$ sachant le jeu courant de paramètres θ^c, et une étape de maximisation (M-step) de cette quantité $Q(\theta,\theta^c)$ par le choix d'un nouvel ensemble de paramètres θ. Les algorithmes GEM diffèrent des algorithmes EM par le fait que les étapes de maximisation se contentent de faire augmenter Q. Cela suffit à garantir la convergence de la vraisemblance vers un maximum local [40].

Pour chaque source j et point temps-fréquence (f,n), l'algorithme [136] introduit R_j variables aléatoires indépendantes gaussiennes $s_{jr,fn}$ $(r = 1,...,R_j)$ appelées sous-sources et distribuées comme $s_{jr,fn} \sim \mathcal{N}_{\mathbb{C}}(0,v_{j,fn})$. Une source de bruit additif isotrope $\mathbf{b}_{fn}$ de covariance diagonale $\Sigma_{\mathbf{b}_{fn}} = \sigma_f^2\mathbf{I}_I \in \mathbb{C}^{I\times I}$ est également ajoutée. Avec ces changements, (3.6) devient :

$$\mathbf{x}_{fn} = \mathbf{A}_f\mathbf{s}_{fn} + \mathbf{b}_{fn} \tag{3.26}$$

où $\mathbf{A}_f \in \mathbb{C}^{I\times R}$ (resp. $\mathbf{s}_{fn} \in \mathbb{C}^R$) résulte de la concaténation $(R = \sum_{j\in\mathcal{J}} R_j)$ des matrices de mélange $\mathbf{A}_{j,f}$ (resp. de toutes les sous-sources $s_{jr,fn}$) de toutes les sources $j \in \mathcal{J}$. Les données observées étant $\mathbf{X} = \left\{\mathbf{x}_{fn}\right\}_{fn}$ et les données non observées $\mathbf{S} = \left\{\mathbf{s}_{fn}\right\}_{fn}$, la quantité Q s'écrit à une constante près :

$$\begin{aligned} Q(\theta,\theta^c) \stackrel{c}{=} -\sum_{fn} \frac{1}{\sigma_f^2}\mathrm{tr}\left[\widehat{\mathbf{R}}_{\mathbf{x}_{fn}} - \mathbf{A}_f\widehat{\mathbf{R}}_{\mathbf{xs}_{fn}}^H - \widehat{\mathbf{R}}_{\mathbf{xs}_{fn}}{\mathbf{A}_f}^H + \mathbf{A}_f\widehat{\mathbf{R}}_{\mathbf{s}_{fn}}{\mathbf{A}_f}^H\right] \\ - \sum_{j\in\mathcal{J},fn} R_j d_{IS}(\widehat{\xi}_{j,fn}|v_{j,fn}), \end{aligned} \tag{3.27}$$

avec : $\widehat{\mathbf{R}}_{\mathbf{x}_{fn}} = \mathbb{E}[\mathbf{x}_{fn}{\mathbf{x}_{fn}}^H]$, $\widehat{\mathbf{R}}_{\mathbf{xs}_{fn}} \stackrel{\Delta}{=} \mathbb{E}[\mathbf{x}_{fn}{\mathbf{s}_{fn}}^H|\theta^c]$, $\widehat{\mathbf{R}}_{\mathbf{s}_{fn}} \stackrel{\Delta}{=} \mathbb{E}[\mathbf{s}_{fn}{\mathbf{s}_{fn}}^H|\theta^c]$ et $\widehat{\xi}_{j,fn} \stackrel{\Delta}{=} \frac{1}{R_j}\sum_{r=1}^{R_j}\mathbb{E}\left[|s_{jr,fn}|^2|\theta^c\right]$. En partant de (3.27), il est démontré que $\left\{\widehat{\mathbf{R}}_{\mathbf{x}_{fn}},\widehat{\mathbf{R}}_{\mathbf{xs}_{fn}},\widehat{\mathbf{R}}_{\mathbf{s}_{fn}}\right\}_{f,n}$ est un ensemble de statistiques suffisantes [134] de $\mathbf{Z}$. Ceci conduit aux deux étapes suivantes de l'algorithme GEM.

E-step Cette étape consiste à calculer les statistiques suffisantes sachant θ^c et $\mathbf{X}$:

$$\widehat{\mathbf{R}}_{\mathbf{s}_{fn}} = \mathbf{\Omega}_{\mathbf{s}_{fn}}\widehat{\mathbf{R}}_{\mathbf{x}_{fn}}{\mathbf{\Omega}_{\mathbf{s}_{fn}}}^H + (\mathbf{I}_R - \mathbf{\Omega}_{\mathbf{s}_{fn}}\mathbf{A}_f)\mathbf{\Sigma}_{\mathbf{s}_{fn}} \in \mathbb{C}^{R\times R} \tag{3.28}$$

$$\widehat{\mathbf{R}}_{\mathbf{xs}_{fn}} = \widehat{\mathbf{R}}_{\mathbf{x}_{fn}}{\mathbf{\Omega}_{\mathbf{s}_{fn}}}^H \in \mathbb{C}^{I\times R} \tag{3.29}$$

avec :

$$\mathbf{\Sigma}_{\mathbf{s}_{fn}} = diag([\phi_{r,fn}]_{r=1}^R) \in \mathbb{R}_+^{R\times R} \tag{3.30}$$

$$\mathbf{\Sigma}_{\mathbf{x}_{fn}} = \mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s}_{fn}}{\mathbf{A}_f}^H + \mathbf{\Sigma}_{\mathbf{b}_{fn}} \in \mathbb{C}^{I\times I} \tag{3.31}$$

$$\mathbf{\Omega}_{\mathbf{s}_{fn}} = \mathbf{\Sigma}_{\mathbf{s}_{fn}}{\mathbf{A}_f}^H{\mathbf{\Sigma}_{\mathbf{x}_{fn}}}^{-1} \in \mathbb{C}^{R\times I} \tag{3.32}$$

où $\phi_{r,fn} = v_{j,fn}$ si $r \in \mathcal{R}_j$ (c'est-à-dire, r est une sous-source de j).

M-step Dans cette étape, les paramètres qui composent l'ensemble θ sont mis à jour de sorte à faire croitre la quantité Q en annulant sa dérivée par rapport à $\mathbf{A}_f$ [136]. La mise à jour des paramètres spatiaux est

$$\mathbf{A}_f = \left[\sum_n \widehat{\mathbf{R}}_{\mathbf{xs}_{fn}}\right]\left[\sum_n \widehat{\mathbf{R}}_{\mathbf{s}_{fn}}\right]^{-1}. \tag{3.33}$$

Les paramètres spectraux sont mis à jour suivant (3.24) afin de minimiser

$$\sum_{j\in\mathcal{J},fn} R_j d_{IS}(\widehat{\xi}_{j,fn}|v_{j,fn}) = \sum_{j\in\mathcal{J}} R_j D_{IS}(\widehat{\Xi}_j|V_j) \tag{3.34}$$

avec $\widehat{\Xi}_j = [\widehat{\xi}_{j,fn}]_{fn} \in \mathbb{R}_+^{F\times N}$ et $\widehat{\xi}_{j,fn} = \frac{1}{R_j}\sum_{r=1}^{R_j} \widehat{\mathbf{R}}_{\mathbf{s}_{fn}}(r,r)$. Pour le modèle $V_j = W_j H_j$, la mise à jour (3.24) devient alors :

$$W_j \leftarrow W_j \odot \frac{(V_j^{\cdot[-2]} \odot \widehat{\Xi}_j)H_j^T}{V_j^{\cdot[-1]}H_j^T}. \tag{3.35}$$

Nous verrons plus tard dans la partie 6.1 des modèles spectraux plus élaborés compatibles avec cet algorithme, et la partie 7.2 décrira une variante avec un second paramètre spatial.

3.3.3 Contraintes

Une des principales raisons du succès inter-disciplinaire de la NMF est sa capacité intrinsèque à représenter une observation comme une somme de composantes. Cependant si on considère en l'état les deux algorithmes de la partie précédente, rien ne garantit que les composantes qui en résulteront auront une quelconque signification [20]. On pourra parfois obtenir de bons résultats de séparation mais sans avoir pu exploiter le fait que les composantes représentent des éléments concrets. Par exemple dans le cas de la musique, on pourrait s'attendre à ce que les composantes correspondent à des notes et exploiter cet aspect de la NMF en ajoutant des contraintes de plus haut niveau (par exemple de structure musicale). Il faut donc trouver des moyens d'ajouter des contraintes à l'estimation de ces composantes.

À l'estimation L'insertion de contraintes peut se faire par ajout de termes de pénalité à la fonction de coût utilisée [93] ou encore par la prise en compte d'à priori [41, 72].

Dans tous les cas, les contraintes que l'on cherche à imposer peuvent concerner :

- la continuité des activations temporelles H [21, 153, 177, 178] (*temporal smoothness*),
- la continuité des composantes spectrales W [172, 177] (*spectral smoothness*),
- l'harmonicité des composantes spectrales W [69, 83, 172],
- la parcimonie de W et H [93, 178].

Chaque contrainte cherche alors explicitement à produire une solution plus réaliste ou du moins probable de la NMF.

Modèles avec structures temporelles Une fois que l'on a une meilleure certitude sur la signification des composantes, l'ajout d'un modèle supplémentaire décrivant l'évolution temporelle des activations (par exemple pour représenter une mélodie plus probable) ou encore leurs dépendances à un instant donné (par exemple pour représenter un accord) est alors possible. On peut citer comme exemple les travaux décrits dans [72,125,127,132,169]. Dans cette situation, l'utilisation de HMM est alors naturelle.

Modèle excitation-filtre Le concept de modèle excitation-filtre [1] a initialement été créé pour modéliser le système humain de production de la parole [62] avant d'être repris pour la modélisation des instruments de musique [83] et pour les modèles de sources en séparation [48, 49, 105, 106, 136].

Ce concept repose sur la dissociation entre ce qui produit le son (l'excitation) et ce qui propage le son (le filtre). Lorsque l'on transpose ce concept dans le domaine temps-fréquence accompagné d'une décomposition NMF, cela donne le modèle de source (ou de spectrogramme) suivant :

$$V_j = V_j^e \odot V_j^\phi = W_j^e H_j^e \odot W_j^\phi H_j^\phi, \tag{3.36}$$

où l'exposant ^{e} est associé à l'excitation et l'exposant $^\phi$ est associé au filtre. Les quatre matrices résultant de la NMF (3.36) sont comme suit :

- $W_j^e \in \mathbb{R}_+^{F \times D^e}$ est un dictionnaire spectral pouvant être un ensemble de spectres inharmoniques, harmoniques et/ou à large bande [136]. Un tel dictionnaire peut être appris sur des données d'apprentissage ou bien estimé à partir des mélanges.
- $H_j^e \in \mathbb{R}_+^{D^e \times N}$ regroupe les activations temporelles correspondantes qui encodent par exemple une partition de musique sous la forme d'un *piano-roll* [60, 71, 95], ou d'une succession de fréquences fondamentales [50].
- $W_j^\phi \in \mathbb{R}_+^{F \times D^\phi}$ est un dictionnaire d'enveloppes spectrales par exemple associé à différents phonèmes dans le cas de la parole [105, 106] et à différents modes de jeu (sourdine) dans le cas d'un instrument de musique [154].
- $H_j^\phi \in \mathbb{R}_+^{D^\phi \times N}$ regroupe les activations temporelles correspondantes qui encodent par exemple une séquence de phonèmes ou des changements de mode de jeu.

En ce qui concerne les mises à jour multiplicatives, (3.24) devient avec ce nouveau modèle :

$$W_j^\phi \leftarrow W_j^\phi \odot \frac{(V_j^e \odot V_j^{\cdot[\beta-2]} \odot X)[H_j^\phi]^T}{(V_j^e \odot V_j^{\cdot[\beta-1]})[H_j^\phi]^T}. \tag{3.37}$$

Contraintes sous forme de signaux de référence Enfin, une autre façon de contraindre la décomposition NMF est d'utiliser un signal de référence qui partage certaines propriétés de sa décomposition avec une des sources à estimer. Le chapitre 6 s'intéresse particulièrement à ces aspects et propose des avancées dans ce domaine. Je donne ci-après un état de l'art des techniques de séparation guidée, notamment par signal de référence, qui ne se retreint pas aux décompositions NMF.

1. Pour éviter les confusions, je préfère parler d'excitation-filtre plutôt que de source-filtre même si ce dernier terme est le plus répandu dans la littérature.

3.4 Séparation guidée

Qu'il s'agisse de modèles NMF ou pas, plus les modèles de sources et les algorithmes de séparation incorporent d'information sur les sources plus ils ont de chances d'être performants. Ces approches sont regroupées sous le nom de séparation de sources guidée par opposition aux approches aveugles et totalement informées (voir partie 4.1).

La volonté d'incorporer dans les algorithmes de séparation de l'information extérieure remonte à la genèse des approches de séparation sous-déterminée, c'est-à-dire au début des années 2000. Il s'agissait alors d'information sur le comportement général des sources et/ou les conditions acoustiques d'enregistrement, par exemple concernant la parcimonie des sources ou le temps de réverbération de la pièce. Ces approches peuvent être qualifiées de faiblement guidées (*weakly guided* [173], voir aussi la discussion de la partie 3.2). Des approches fortement guidées [155, 183] (*strongly guided* [173]) ont ensuite fait leur apparition, avant qu'elles ne soient explorées de façon intensive ces cinq dernières années. C'est sur cette dernière catégorie d'approches que porte cette partie et se focalise cette thèse.

La spécificité des approches fortement guidées est d'utiliser des informations spécifiques à un signal (valables à un instant) et non à une classe de signaux. Le guidage peut concerner aussi bien les aspects spatiaux que spectraux, mais je me concentrerai ici uniquement sur les approches de séparation guidées au niveau spectral. Je présenterai en particulier des approches de séparation guidée par signal de référence qui est le cas particulier que j'explorerai dans cette thèse.

3.4.1 Cartographie et classification des approches fortement guidées

Liutkus *et al.* [111] ont proposé en 2013 un tour d'horizon des approches guidées et informées. Le nombre de travaux sur les aspects guidés s'est depuis encore agrandi, il devient maintenant intéressant d'en établir une classification plus précise et rigoureuse, chose qui n'avait jamais été faite auparavant. La cartographie de ces approches sera d'autant plus utile pour positionner l'approche générale proposée dans le chapitre 6 et qui concerne les approches guidées par signal de référence.

Je propose donc la classification des approches fortement guidées selon les trois critères cumulables suivants :

— **intervention d'un utilisateur** : Les approches qui sont ici qualifiées de *user-guided* sont celles où l'utilisateur produit de nouvelles informations avec comme intention d'améliorer la séparation. Cet ensemble d'approches inclut également les approches qui réutilisent des informations qui sont déjà existantes et que l'utilisateur met juste en correspondance avec les sources. De plus, les approches faiblement guidées impliquant la formulation d'hypothèses à propos des sources ou du mélange (*cf.* partie 3.2) sont exclues de cet ensemble.

— **utilisation d'informations symboliques** :
Il peut s'agir de transcriptions symboliques (textes, partitions de musique) exploitées telles quelles ou après synthèse, ou encore de renseignements sur des zones temps-fréquence (annotations, sélections). Cet ensemble d'approches re-

groupe celles qui utilisent en entrée des informations symboliques qu'elles soient produites par un utilisateur ou déjà disponibles.

— **utilisation d'un signal de référence** : Un signal de référence est un signal qui comporte des similitudes avec une ou plusieurs sources du mélange à traiter. Il peut être produit par un utilisateur ou par synthèse ou être déjà disponible.

La figure 3.1 regroupe les approches les plus célèbres selon cette classification. Une version étendue est donnée en Annexe A incluant les références bibliographiques. On peut entre autres y voir apparaitre un nombre important d'approches aux intersections de ces trois critères.

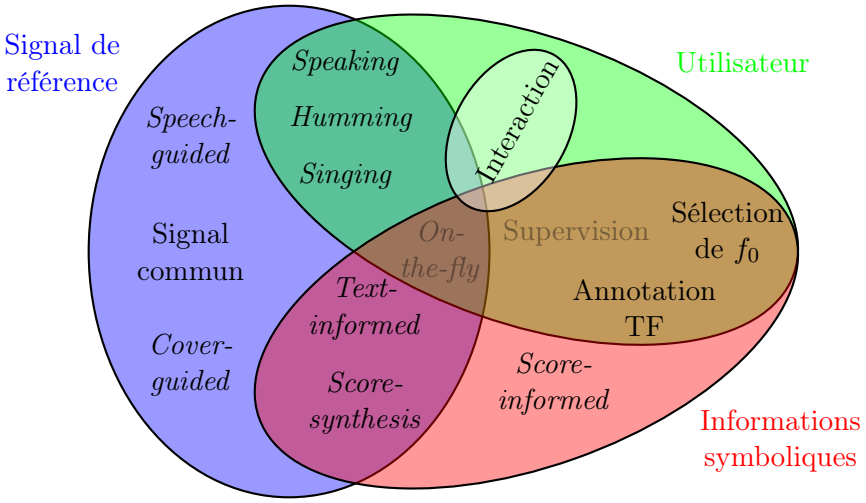


Figure 3.1 – Classification des techniques de séparation de sources fortement guidées.

3.4.2 Quelques approches sans signal de référence

Plusieurs approches utilisent des informations autres qu'un signal de référence.

Score-informed Le terme *score-informed source separation* a fait son apparition en 2006 [183] avant de réapparaitre de façon massive entre 2010 et aujourd'hui. Il s'agit d'exploiter les partitions de musique ou les fichiers MIDI en les utilisant par exemple directement pour contraindre un modèle [59, 82, 87, 95, 154], typiquement par le biais de la fréquence fondamentale f_0. Une autre façon d'exploiter ces informations symboliques est de synthétiser des exemples audio [71, 74] (*score synthesis based separation*). Un large panorama de toutes ces approches est donné par Ewert *et al.* [60].

Annotation de spectrogramme Ce genre d'annotations étant très spécifiques, elles proviennent nécessairement d'un utilisateur cherchant à améliorer la séparation. L'utilisateur opère généralement la sélection de nouvelles zones temps-fréquence [108] ou

d'instants [131] pour définir où les sources sont actives. De la même façon, il peut définir la courbe approximative de f_0 d'une source [50]. Des informations plus précises à propos de ces zones peuvent aussi être précisées, par exemple à propos de la qualité de séparation après chaque itération [44]. Dans le même esprit que les approches précitées utilisant des partitions, l'utilisateur peut avoir à sélectionner des notes de musique pré-délimitées [73] pour définir les sources.

Interaction utilisateur/algorithme Il s'agit d'un nouveau genre d'approches qui propose à un utilisateur expert d'interagir avec l'algorithme de séparation au cours de l'estimation des sources [26, 27, 44]. Replacer l'audition aiguisée d'un utilisateur expert au centre d'un algorithme de séparation est sûrement un bon choix pour obtenir des résultats exploitables, cependant ce processus itératif peut s'avérer fastidieux.

Supervision (faiblement guidée) Le terme supervisé pouvant prendre plusieurs sens, je précise qu'ici il fera référence à la signification que lui donnent les techniques d'apprentissage automatique, c'est-à-dire l'utilisation d'une base de donnée annotée en vue d'une phase d'apprentissage préalable à l'étape de séparation. Cela permet par exemple l'apprentissage de modèles de sources [15, 80, 135]. L'utilisateur sera typiquement amené à mettre en correspondance des informations ou des signaux déjà existants et qui ne seront pas utilisés pendant la séparation. C'est pourquoi, je ne considère pas ces approches comme utilisant un signal de référence.

3.4.3 Guidage par signal de référence

Focalisons-nous maintenant sur les approches exploitant un signal de référence.

3.4.3.1 Signal déjà existant

Cover-guided L'objectif de ces approches [75, 162] est la séparation des instruments d'un morceau de musique original par l'utilisation d'une reprise (*cover song*) dont on détient les pistes séparées. Chacune de ces pistes sert alors de signal de référence pour la séparation du morceau original.

Séparation de signaux communs On parle de signal commun (*common signal*) ou de « séparation de composantes communes » [116] lorsque la même source apparait de façon identique dans deux mélanges différents. Par exemple dans le cas des films en différentes langues, la même musique [2] apparaitra dans les différentes versions mais mélangée avec des dialogues différents [29, 109, 113]. On peut également se retrouver dans cette situation lorsque les références sont des éléments répétés provenant du même film [160, 161]. Le cas d'un enregistrement multicanal réel peut aussi s'y apparenter, notamment lorsque la présence d'une source dans un des canaux peut être négligée, par exemple lorsqu'un micro est proche d'une source et très éloigné d'une autre source. La partie 7.1.1 donne plus de détails sur les techniques qui traitent ce problème.

2. ainsi que les effets.

On-the-fly (faiblement guidée) El Badawy *et al.* [51, 52] ont récemment revisité la séparation de sources supervisée avec le concept de séparation à la volée (*on-the-fly*). Il s'agit pour l'utilisateur de renseigner un mot-clef par source (« vent », « oiseau »), ce qui va permettre de regrouper un certain nombre d'exemples audio correspondants en vue de la séparation. Même si il ne s'agit que d'une simple mise en correspondance du point de vue de l'utilisateur, on peut considérer cette approche comme guidée par signal de référence. En effet, il n'y a pas d'apprentissage au préalable (à l'inverse des méthodes supervisées traditionnelles) et les signaux sont utilisés durant la séparation.

3.4.3.2 Signal produit par l'utilisateur

Les travaux de Smaragdis et Mysore [157] sur la séparation guidée par le fredonnement d'un utilisateur (*separation by humming*) ont ouvert la voix à une série d'approches où l'utilisateur produit lui-même un signal de référence. Initialement, il lui était demandé de fredonner voire d'imiter la voix chantée ou l'instrument à extraire. Le concept a ensuite été repris [67, 85, 105, 106, 160, 180] notamment par des approches où la même phrase est répétée par l'utilisateur.

3.4.3.3 Signal synthétisé

La synthèse de signal à partir d'informations symboliques (texte, partition de musique) permet d'obtenir des signaux de référence exploitables. On peut rappeler l'existence des approches *score synthesis based* [71, 74] et noter l'existence des approches *text-informed* [105, 106] dans lesquelles la transcription phonétique de la source de parole est disponible et permet la synthèse.

3.4.4 Limitation des techniques avec signal de référence

Les approches existantes basées sur un signal de référence ont plusieurs limitations.

Une première observation est que chaque approche vise une situation particulière et/ou un type de source (parole, instrument). L'expertise de l'utilisateur sera dans la plupart des cas indispensable [3] pour choisir l'outil de séparation adapté aux signaux qu'il a à sa disposition.

Deuxièmement, si les références ne sont pas déjà disponibles, l'utilisateur devra soit :

— produire des références. Dans cette situation les conditions d'enregistrement, la qualité de l'imitation sont peu contrôlées ce qui ne garantit pas la qualité finale de séparation.

— rechercher des références, ce qui nécessite du temps supplémentaire ou bien l'utilisation d'outils de recherche.

3. L'approche de séparation à la volée est sans doute la seule à pouvoir s'adresser au grand public.

Chapitre 4

Travaux similaires ou connexes

L'approche SPORES que je propose dans cette thèse est une approche de séparation de sources guidée par signaux de référence où les références sont obtenues par détection de motifs. Le succès de cette approche repose sur l'hypothèse que les sources présentes dans le mélange se répètent soit ailleurs dans le mélange, soit dans un autre contenu.

Ce dernier chapitre de l'état de l'art va permettre de compléter le positionnement de cette nouvelle approche par rapport à deux autres techniques. En effet les deux chapitres précédents se sont focalisés sur les aspects spécifiques à la détection ou à la séparation présents dans SPORES, alors que ce chapitre positionne l'approche SPORES dans son ensemble.

Je présenterai tout d'abord les approches de séparation informée, puis un groupe d'approches exploitant comme SPORES les redondances internes au mélange. Enfin, je décrirai les aspects partagés avec l'approche SPORES ainsi que les originalités apportées par cette thèse.

4.1 Séparation informée ou codage spatial

Les méthodes de séparation de sources informée [110, 133, 137] ou de codage spatial [25, 35, 89] reposent sur l'utilisation d'informations détaillées calculées en amont sur les sources d'origine. On va dans un second temps chercher à réduire la taille de ces informations en vue de leur transmission en parallèle du mélange. L'étape de décodage s'apparente alors à une étape de séparation dans le cas oracle, c'est-à-dire où les modèles des sources sont estimés sur la vérité terrain. Ainsi, ce problème peut s'apparenter à un problème de codage.

Toutes ces méthodes donnent des résultats bien supérieurs à ceux des autres méthodes de séparation, mais ne sont pas applicables dans les scénarios où les sources d'origine ne sont jamais observées. Leur champ applicatif est donc restreint à la transmission efficace des sources.

Principe La principale spécificité des méthodes de séparation de sources informée est la connaissance des sources d'origine. Les paramètres de mélange et des sources sont

alors déterminés empiriquement au préalable en observant le mélange et les sources d'origine (encodage). Ils peuvent être de nature spatiale ou spectrale et peuvent être ensuite réutilisés au moment de la séparation (décodage).

Codage spatial Dans le cas des méthodes de codage spatial (*Spatial Audio Object Coding* ou SAOC) [25, 35, 89], les paramètres de mélange sont choisis à l'encodage pour faciliter l'étape de séparation à venir. Typiquement, les sources sont artificiellement positionnées le plus loin possible les unes des autres. L'étape de séparation utilise alors des informations intercanales comme les différences d'intensité et la cohérence intercanale.

Séparation informée Dans le cas de la séparation de sources informée [110, 133, 137], on peut ne pas avoir le contrôle des paramètres de mélange, par exemple si on veut traiter un mélange « commercial » produit par un ingénieur du son. La transmission de paramètres spectraux permet alors de traiter des mélanges plus réalistes que les mélanges instantanés, convolutifs ou diffus [110]. Un autre avantage est de pouvoir envisager le cas mono-canal contrairement aux méthodes de codage spatial.

Différents modèles spectraux de sources ont déjà été utilisés dans ce cadre, par exemple les décompositions parcimonieuses [138] ou la NMF [133]. Ces modèles permettent intrinsèquement de réduire le nombre de paramètres qui représentent les spectrogrammes des sources, et donc de réduire le débit de transmission.

Encodage des paramètres spectraux et performance La quantification des paramètres spectraux à transmettre peut être effectuée de façon assez efficace, étant donnée leur redondance. Il en résulte un débit de transmission relativement faible (2 kbit/seconde par source [110]) comparé à la transmission des signaux de chaque piste d'origine [1]. La transmission des paramètres peut ensuite se faire directement ou par le biais de méthodes de tatouage audio (*watermarking*) [138, 139].

Les performances d'un tel système sont limitées par les performances oracle du modèle choisi qui est une borne supérieure de distorsion pour le système correspondant. Si l'on décide de transmettre avec pertes les paramètres des sources, alors les performances du système seront inférieures à celles de l'oracle du modèle choisi.

On peut aussi noter qu'il est possible de dépasser cette borne en transmettant une seconde information au décodeur : le résidu. Il s'agit des erreurs commises par l'encodeur et qu'il est possible de mesurer à cette étape en comparant les sorties du décodeur à la vérité terrain. Ces résidus peuvent alors eux aussi être encodés de façon traditionnelle (typiquement par un codeur AAC) et réinsérés au moment du véritable décodage.

Un seconde possibilité récemment proposée par Ozerov *et al.* [133] consiste à utiliser les informations à propos de ces résidus pour encoder directement les sources. On parle alors de séparation informée par codage. Cette formalisation permet notamment de contrôler de façon optimale la distorsion en fonction du débit et ainsi d'atteindre des performances nettement supérieures moyennant une augmentation du débit.

1. La transmission d'une piste ou d'un mélange requiert généralement un débit allant de 32 à 128 kbit/seconde.

Application Le codage spatial a pour principal but de réduire le nombre de canaux transmis et en conséquence le débit global. Il est adapté à toutes les situations où on cherche à transmettre un nombre important d'objets audio avec un débit réduit.

Pour la séparation informée, le scénario d'utilisation le plus fréquent est celui où le morceau de musique a déjà été transmis à un utilisateur et celui-ci décide après coup d'accéder aux pistes séparées de ce morceau. La séparation informée permet alors la transmission à moindre coût des différentes pistes du morceau de musique. L'utilisateur peut ensuite réutiliser les pistes par exemple pour ses propres créations musicales. Les informations utiles à la séparation peuvent par exemple être générées au moment où l'ingénieur du son réalise le mélange en studio.

Certaines techniques engendrent une perte en qualité qui est acceptable pour le grand public. En revanche, un professionnel du domaine de l'audio voudra se procurer la version des pistes séparées ayant la meilleure qualité possible. Il faudra alors utiliser les techniques permettant de dépasser les bornes oracle par l'augmentation du débit, voire transmettre les pistes originales.

4.2 *REPET, REPET-SIM, KAM*

Un autre groupe d'approches issues des travaux de Rafii *et al.* [144–146] exploitent comme SPORES la redondance des signaux pour leur séparation avec comme principal application les morceaux de musique. L'idée commune est l'utilisation d'un filtre médian sur un ensemble de point temps-fréquence pour définir le modèle d'une source qui se répète. Ce paragraphe se limite à la description des approches suivantes :

— *REPET* [144, 146] (*REpeatting Pattern Extraction Technique*)
— *REPET-SIM* [145] (*REPET using the SImilarity Matrix*)
— *KAM* [112] (*Kernel Additive Models*)

Les approches *REPET* [144, 146] et *REPET-SIM* [145] sont des approches aveugles de séparation voix/musique (deux sources) qui supposent que le fond musical se répète au cours du temps alors que la voix ne se répète pas. Dans *REPET* le fond musical est supposé périodique alors que *REPET-SIM* autorise n'importe quelle répétition au cours du temps. L'approche *KAM* [112] est une généralisation à un nombre quelconque de sources qui autorise n'importe quelle structure de répétition en temps et en fréquence pour chaque source.

Le point commun de toutes ces approches est la succession des étapes algorithmiques :

— identification des répétitions,
— modélisation des segments répétés,
— filtrage des motifs répétés.

Ces trois étapes sont décrites ci-après pour chaque approche.

4.2.1 Identification des répétitions

Le but de cette étape est de constituer pour chaque point temps-fréquence (f,n) un ensemble E_{fn} de points temps-fréquence présentant une forte similarité. Cette similarité peut être identifiée à l'échelle de la trame par exemple en supposant que le fond musical se répète périodiquement [144, 146].

Hypothèse de répétition périodique des trames (*REPET* [144, 146]) On cherche à déterminer la période de répétition p la plus probable et ainsi en déduire l'ensemble des trames répétées pour construire l'ensemble des points temps-fréquence tel que :

$$\mathsf{E}_{fn} = \{(f,n+kp)\}_{\forall k|(n+kp)\in[1,N]} \tag{4.1}$$

La détermination de p peut se faire par l'analyse d'un *beat spectrum* [70] comme effectuée dans [144, 146].

Hypothèse de répétition intermittente des trames (*REPET-SIM* [145]) Les trames répétées sont identifiées par le biais d'une matrice de similarité S (2.9) du mélange x. Chaque élément $S(n,n')$ est alors la mesure de similarité cosinus (2.8) entre les deux trames n et n' de la STFT [2]. L'ensemble E_{fn} est alors constitué des paires (f,n') pour lesquelles $S(n,n')$ est la plus élevée. En d'autres termes, un nombre prédéfini de trames n' présentant le plus de similarité avec la trame n sont sélectionnées.

Hypothèse de noyau de proximité (*KAM* [112]) Dans ce cas, l'ensemble E^{j}_{fn} est défini pour chaque source j et peut inclure des points temps-fréquence quelconques. On appelle cet ensemble « noyau de proximité » en raison de la proximité des points temps-fréquence qui le composent. Par exemple une source de type percussif aura un noyau composé des éléments proches verticaux :

$$\mathsf{E}^{\updownarrow}_{fn} = \{(f+k,n)\}_{\forall k\in[-K,\dots,K]} \tag{4.2}$$

où K est la hauteur du noyau. Une source comportant des harmoniques stables aura un noyau composé des éléments proches horizontaux :

$$\mathsf{E}^{\leftrightarrow}_{fn} = \{(f,n+k)\}_{\forall k\in[-K,\dots,K]} \tag{4.3}$$

où K est la largeur du noyau. Une source de voix aura un noyau « neutre » composé des éléments proches horizontaux et verticaux :

$$\mathsf{E}^{\text{voix}}_{fn} = \mathsf{E}^{\updownarrow}_{fn} \cup \mathsf{E}^{\leftrightarrow}_{fn}. \tag{4.4}$$

On peut aussi construire des noyaux combinant une structure de répétition (4.1) avec des structures locales (4.2), (4.3), (4.4). Les noyaux de proximité (*KAM*) généralisent les approches *REPET* et *REPET-SIM*.

2. en amplitude

4.2.2 Modélisation du segment répété

Une fois les ensembles E_{fn}^{j} identifiés ou choisi, ils sont exploités dans le but de construire des modèles $v_{j,fn}$ du spectre de puissance des sources j correspondantes.

Moyenne géométrique La première approche [144] a été l'utilisation d'une moyenne géométrique uniquement pour le fond musical

$$v_{fn} = \left(\prod_{(f',n') \in \mathsf{E}_{fn}} |x_{f'n'}|^2 \right)^{\frac{1}{\#\mathsf{E}_{fn}}} \tag{4.5}$$

suivie d'un seuillage [144].

Médiane Le filtre médian a par la suite été préféré pour les autres approches [112,145, 146]. Pour les approches de séparation voix/musique [145, 146], uniquement les zones temps-fréquence du fond musical sont modélisées :

$$v_{fn} = \underset{(f',n') \in \mathsf{E}_{fn}}{\text{médiane}} \{|x_{f'n'}|^2\}. \tag{4.6}$$

L'utilisation d'une médiane se justifie par la parcimonie et la variabilité de la représentation temps-fréquence de la source de voix. Par exemple pour une période p donnée, les points temps-fréquence avec une forte variation seront exclus de l'estimation par le filtre médian alors qu'ils auraient à l'inverse affecté la moyenne.

De la même façon, l'approche multi-source [112] utilise un filtre médian pour chaque source j :

$$v_{j,fn} = \underset{(f',n') \in \mathsf{E}_{fn}^{j}}{\text{médiane}} \{|x_{f'n'}|^2\}. \tag{4.7}$$

4.2.3 Filtrage

Le filtrage est effectué dans le domaine temps-fréquence par l'application d'un gain au mélange (voir équation (3.13)) et par reconstruction dans le domaine temporel par ISTFT.

Différents types de masques ont été utilisés :

— des masques binaires (3.14) pour la première approche [144],

— un masque spécifique pour le fond musical

$$G_{fn} = \frac{\min\{v_{fn}, |x_{fn}|^2\}}{|x_{fn}|^2} \in [0,1] \tag{4.8}$$

pour les approches qui ont suivi [145, 146] [3],

— un gain de Wiener (3.15) pour les différentes sources dans le cas de *KAM*. Dans ce dernier cas, la source de voix doit être modélisée pour être séparée par filtrage de Wiener et ne peut pas être obtenue comme un résidu par simple soustraction.

3. Pour ces deux approches, le résidu qui correspondant à la voix est obtenu par soustraction du signal de musique estimé dans le domaine temporel.

4.3 Lien avec l'approche SPORES

L'approche SPORES se place dans le cas où les sources ne sont jamais observées, c'est-à-dire le cas « non informé ». Cependant pour ne pas confondre approches informées et guidées qui sont des termes pouvant être perçu comme synonymes, il était important de bien décrire ces différences même si elles sont marquées.

D'autre part, des parentés importantes existent entre les approches de type REPET présentées dans la partie précédente et l'approche SPORES développée dans cette thèse. Toutefois, la différence majeure est le fait que SPORES permet de faire des hypothèses plus fortes que de simples répétitions dans le domaine temps-fréquence par la prise en compte de déformations.

Observation des sources originelles Le cas « informé », c'est-à-dire quand on détient la vérité terrain avant que le mélange soit généré, n'est pas pris en compte par l'approche SPORES. Ce cas ne faisant pas partie du champ applicatif de SPORES, si les sources sont disponibles, elles seront directement transmises et le problème de la séparation ne se pose pas. En effet, les applications ciblées par notre partenaire industriel nécessitent des sources séparées de la meilleure qualité possible (voir chapitre 8).

En revanche, SPORES inclut le cas où une des sources est observée conjointement à d'autres signaux sans intérêt [109, 113], c'est-à-dire au sein d'un autre mélange que celui que l'on cherche à séparer. Cette situation est généralement appelée séparation de signaux communs (*common signal separation*) [111]. Le signal de référence est alors un mélange. On peut inclure dans ce type de situation le cas où la référence contient un *sample* de la source originelle. *REPET* peut aussi s'apparenter à cette situation puisqu'il cherche à utiliser les répétitions du fond musical sans modéliser la déformation entre les répétitions.

Dimension des *patterns* L'approche SPORES prend en compte des signaux de référence de l'ordre de quelques secondes à plusieurs minutes. Les ressemblances entre trames sont donc exploitées uniquement dans le contexte de l'appariement d'un ensemble important de trames. C'est aussi le cas de l'approche originelle REPET où le modèle du fond musical se construit sur la base d'une période de répétition, ainsi le fond musical est modélisé comme une succession de séquences identiques.

En revanche, *REPET-SIM* et *KAM* vont utiliser la ressemblance entre des trames une à une voire entre des points temps-fréquence. On peut alors parler de modèles à échelle locale par opposition à des modèles reposant sur le long terme comme SPORES ou REPET.

Deuxième partie

Contributions

Chapitre 5

Détection robuste de motifs

De façon générale, les difficultés rencontrées au cours de la tâche de détection de motifs sont liées soit à la présence d'autres sources (ou de bruit), soit aux déformations entre occurrences du même motif. L'étude de techniques détection robuste à la présence d'autres sources a été ici préférée au cas des motifs déformés. D'une part parce qu'il était préférable de répondre séparément au cas des motifs déformés et au cas des motifs présents avec d'autres sources, les effets de ces deux altérations se confondant. Et d'autre part en raison du potentiel des motifs identiques comme référence pour la séparation guidée comme nous le verrons dans le chapitre 7 où des techniques de séparation de signaux communs sont étudiées. Ainsi, la DTW ne sera pas utilisée ici. De plus, afin de faciliter la lecture des résultats, le parcours de la base de recherche a été simplifié pour éviter les « faux-négatifs ».

Dans ce chapitre, je m'attacherai à exploiter le fait que la distorsion due à la présence d'une source étrangère aux motifs cibles est parcimonieuse en temps et en fréquence. Dans ce contexte, des *features* cepstraux tels les MFCC ne permettent pas une détection robuste aux bruit car cette distorsion aurait alors un effet sur l'ensemble des *features*. À l'inverse, les représentations spectrales telles que les chromas ou la STFT conservent la parcimonie.

Je présente ici une étude sur le choix de la distance pour la comparaison de séquences STFT. L'objectif est d'adapter la distance pour la détection de motifs dans la situation où les motifs sont en présence d'autres sources. L'idée est d'exploiter la parcimonie du bruit (les autres sources) en utilisant une distance l_p avec des valeurs faibles de p.

La première partie de l'étude cherche à déterminer empiriquement les valeurs optimales de p pour différents ensembles d'apprentissage représentant les tâches de :

— comparaison de motifs identiques de musique en présence de voix,
— et de comparaison de motifs identiques d'un instrument au sein d'un morceau de musique.

Ces valeurs sont ensuite évaluées expérimentalement sur la tâche de recherche de motifs de musique dans des mélanges voix/musique artificiels. On s'intéressera en particulier au cas où la requête est elle aussi un mélange voix/musique. On cherche ainsi à représenter notre cas d'usage, c'est-à-dire la recherche d'un motif présent dans un

segment de bandes-son de film où plusieurs sources sont présentes, typiquement de la voix et de la musique. La séparation des sources de ce segment est ensuite facilitée par la présence des motifs de musique trouvés qui peuvent être exploités par des approches de séparation guidée par signal de référence comme celles présentées dans les chapitres 6 et 7.

Ces distances ont par ailleurs de nombreuses autres possibilités d'utilisation qui sont ensuite décrites.

5.1 Parcimonie des distances entre mélanges

L'hypothèse de travail de cette partie est que les différences induites par la superposition d'autres sources sont parcimonieuses dans le domaine temps-fréquence. Le but est ici de mesurer la parcimonie des représentations STFT dans le cas de sources qui se superposent afin de l'exploiter pour améliorer la détection de motifs en présence d'autres sources. Je m'intéresse notamment à l'apprentissage des valeurs optimales de p pour différents types de sources.

5.1.1 Données

Dans les différentes situations décrites ci-après, chaque corpus d'apprentissage est divisé en 10 sous-corpus sur lesquels seront effectués 10 apprentissages séparément. Un sous-corpus est composé d'une paire de signaux dont les différences absolues d'amplitude des éléments de leurs STFT sont utilisées dans l'apprentissage de p. Ces deux signaux sont des mélanges de deux types de sources. Le signal d'une des deux sources est identique dans les deux mélanges et les signaux de l'autre type de sources sont différents (voir Figure 5.1).

Mélanges voix/musique Dans ce cas, ce sont les signaux de musique qui sont identiques. Les exemples de musique sont 10 morceaux de musique complets d'une durée moyenne de 4 minutes. Les exemples de voix sont différents et tirés d'une base de données précédemment enregistrée [17]. Les niveaux entre musique et voix sont fixés parmi les valeurs suivantes (en dB) : -12, -6, 0, 6, 12 et $+\infty$ [1]. Ils peuvent être différents d'un mélange à l'autre. Ces valeurs représentent les niveaux usuels observés dans les bandes-son de films entre musique et voix (typiquement -12 et 6 dB) [2].

Morceaux de musique Dans ce cas, ce sont les signaux d'un instrument de musique cible (basse, batterie, guitare ou l'ensemble des instruments [3]) qui sont identiques entre les deux mélanges. Les signaux de l'autre type de source sont alors composés du reste

1. $+\infty$ désigne le cas où la musique est seule.
2. Ces deux rapports musique / voix seront réutilisés dans les chapitres suivants. On parlera alors de rapport voix/musique
3. Dans ce cas, c'est la voix chantée qui compose l'autre type de source. En effet, il est plus pertinent de s'intéresser aux répétitions d'un fond musical plutôt qu'aux répétitions de voix qui sont beaucoup plus variables.

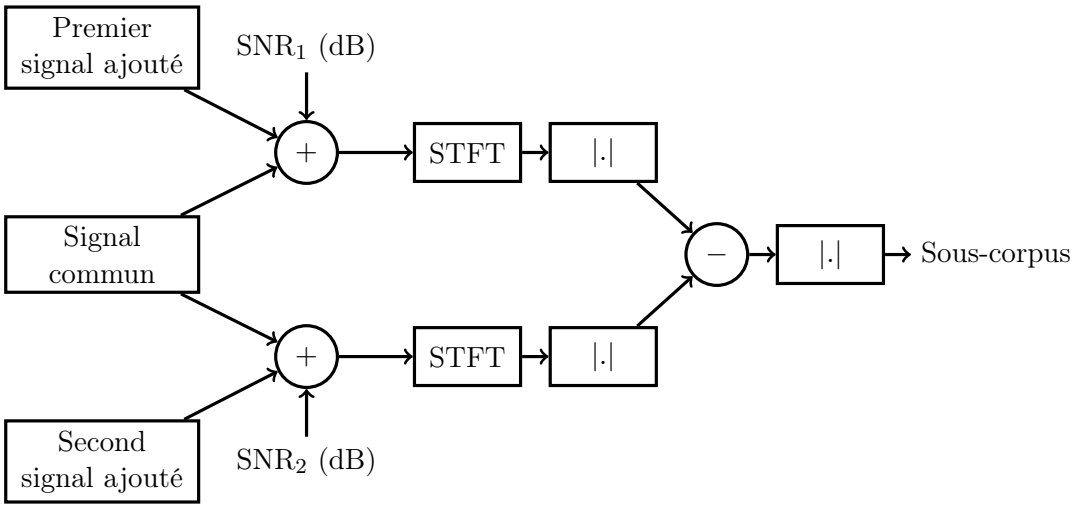


Figure 5.1 – Schéma de génération d'un sous-corpus d'apprentissage à partir de deux signaux du même type et un signal commun d'un autre type.

des instruments au même instant et du reste des instruments à un autre instant dans le morceau. Ce deuxième exemple n'est pas choisi aléatoirement et correspond à un segment où l'instrument cible rejoue les mêmes notes mais les autres instruments jouent des notes différentes. Les 10 extraits sont d'une durée moyenne de 30 secondes.

Les niveaux entre l'instrument cible et le reste des instruments sont fixés parmi les valeurs suivantes (en dB) : -12, -9, -6, -3, 0 et $+\infty$. Ces valeurs représentent les niveaux usuels observés dans les morceaux de musique, typiquement $-7{,}5$ dB pour la basse et la batterie, -10 dB pour la guitare et -4 dB pour la voix.

5.1.2 Apprentissage de p

L'hypothèse de parcimonie des différences entre coefficients des STFT rend naturelle l'utilisation des distances l_p avec $p \leq 2$ qui donnent moins d'importance aux grands écarts. Cependant, la distribution empirique de ces différences présente une asymétrie pour les sous-corpus d'apprentissage ayant des mélanges de niveaux différents. En effet, les différences de signes positifs correspondent au premier signal ajouté et les différences de signes négatifs au second signal ajouté (voir Figure 5.1). Cette asymétrie n'est pour le moment pas prise en compte.

L'apprentissage du paramètre p est effectué par estimation au sens du maximum de vraisemblance des paramètres d'une distribution exponentielle généralisée [170] :

$$P(d_{fn}) = p \times \frac{\beta^{\frac{1}{p}}}{\Gamma(\frac{1}{p})} \times e^{-\beta \times {d_{fn}}^p} \tag{5.1}$$

où d_{fn} est la différence absolue entre *features*, $p \in \mathbb{R}^+$ et $\beta \in \mathbb{R}^+$ sont des paramètres

de la distribution et Γ est la fonction gamma. La log-vraisemblance est alors :

$$\mathcal{L} = -\sum_{fn} \Big(\log(p) + \frac{1}{p}\log(\beta) - \log(\Gamma(\frac{1}{p})) - \beta \times {d_{fn}}^p \Big). \tag{5.2}$$

En pratique, nous considérons les paramètres $\log(p) \in \mathbb{R}$ et $\log(\beta) \in \mathbb{R}$ qui donnent une héssienne mieux conditionnée. Nous utilisons alors la fonction Matlab « fminunc » pour déterminer les valeurs de paramètres $\log(p)$ et $\log(\beta)$ qui maximiser cette log-vraisemblance.

5.1.3 Résultats

Les résultats présentés dans les Tableaux 5.1 et 5.2 sont les moyennes des valeurs de p apprises sur les 10 sous-corpus d'apprentissage. Pour l'ensemble des moyennes affichées, les écarts-types sont inférieurs à 0,1. Les résultats sont regroupés dans le Tableau 5.1 pour l'expérience avec des mélanges voix/musique et dans le Tableau 5.2 pour l'expérience avec des instruments dans des morceaux de musique.

On remarque au travers des différents tableaux que la valeurs de p dépend essentiellement du type de sources présentes dans les mélanges plus que des niveaux de mélange. On observe des valeurs typiques de $p = 0{,}18$ pour les morceaux de musique dans de la voix, $p = 0{,}39$ pour la basse dans un morceau de musique et $p = 0{,}24$ pour un fond musical en présence de voix chantée. Les tableaux pour la batterie et la guitare ont été placés en Annexe D.1. Ils montrent des résultats similaires avec typiquement $p = 0{,}31$ pour la batterie et $p = 0{,}38$ pour la guitare.

	-12 **dB**	-6 **dB**	0 **dB**	6 **dB**	12 **dB**	$+\infty$
-12 **dB**	0,18	0,18	0,17	0,17	0,16	0,15
-6 **dB**	0,18	0,18	0,18	0,18	0,17	0,15
0 **dB**	0,17	0,18	0,18	0,18	0,18	0,15
6 **dB**	0,17	0,18	0,18	0,19	0,18	0,16
12 **dB**	0,16	0,17	0,18	0,19	0,19	0,16

Tableau 5.1 – Valeur moyenne de p entre deux occurrences d'un morceau de musique mélangé à de la voix à deux niveaux différents.

5.2 Détection de répétitions exactes de musique

Dans cette partie, différentes valeurs de p sont testées dans une situation de détection de motifs identiques de musique dans de la parole. Les résultats sont affichés sous la forme de courbe précision-rappel qui permettent de comparer plusieurs systèmes de détection. Les distances l_p sont notamment comparées à la distance cosinus (2.8) qui est plus couramment utilisée [145] pour comparer l'amplitude des STFT.

	−12 **dB**	−9 **dB**	−6 **dB**	−3 **dB**	0 **dB**	$+\infty$
−12 **dB**	0,39	0,39	0,38	0,38	0,38	0,46
−9 **dB**	0,39	0,39	0,39	0,38	0,38	0,46
−6 **dB**	0,39	0,39	0,39	0,39	0,38	0,46
−3 **dB**	0,40	0,39	0,39	0,39	0,39	0,46
0 **dB**	0,40	0,40	0,40	0,40	0,39	0,46

(a) Apprentissage pour la basse.

	−12 **dB**	−9 **dB**	−6 **dB**	−3 **dB**	0 **dB**	$+\infty$
−12 **dB**	0,24	0,24	0,24	0,23	0,23	0,20
−9 **dB**	0,24	0,24	0,24	0,24	0,23	0,20
−6 **dB**	0,24	0,24	0,25	0,24	0,24	0,20
−3 **dB**	0,24	0,24	0,24	0,25	0,25	0,20
0 **dB**	0,23	0,24	0,24	0,25	0,25	0,20

(b) Apprentissage pour le fond musical en présence de voix chantée.

Tableau 5.2 – Valeur moyenne de p entre deux occurrences d'un instrument de musique mélangé à deux niveaux différents dans des portions différentes du morceau avec tous les instruments.

5.2.1 Données

Motifs musicaux Un total de 50 motifs sont tirés des 10 morceaux de musique déjà utilisés dans la partie précédente (5 par morceau). Ce sont des extraits de 4 secondes contenant l'ensemble des instruments. Ils seront utilisés pour composer les requêtes et les segments de la base de recherche contenant les motifs recherchés.

Niveaux des mélanges Que ce soit dans la requête ou dans la base de recherche, les motifs sont mélangés avec de la voix à différents niveaux (en dB) : 12, 0, −3, −6, −9 et −12. Ces niveaux sont globalement plus bas que pour l'apprentissage de p afin de mieux observer les différences entre systèmes de détection. Dans les expériences rapportées par la suite, uniquement des résultats pour un niveau des motifs dans la requête de −12 dB seront affichés afin de mieux reproduire notre cas d'usage.

Bases de recherche Pour un motif donné, 20 copies (motifs recherchés) de celui-ci sont disséminées dans 1 heure de signal de parole provenant de [17]. Les bases de recherche sont différentes pour chaque requête. Elles ne contiennent cependant pas d'autres exemples de musique.

5.2.2 Systèmes de détection

Les systèmes de détection utilisés dans cette expérience ont été simplifiés afin d'annuler l'impact sur les résultats que pourraient avoir des techniques comme la DTW ou

la recherche par graine (formalisme ARGOS [88]). La totalité de la requête est comparée à tous les segments possibles de même taille dans la base de recherche.

Chaque comparaison produit un score qui est la somme des distances entre trames (voir équation (2.7)). Les comparaisons sont effectuées sur les amplitudes de la STFT par une distance l_p ou cosinus (voir équations (2.7) et (2.8)). La réponse d'un système de détection est constituée des N segments ayant le plus petit score (N étant un paramètre variable du système). Ainsi, ces systèmes ont deux paramètres : le paramètre p de la distance utilisée et le nombre N de motifs de sortie.

5.2.3 Résultats

Génération des courbes précision-rappel La précision pour un système donné et une requête donnée est le rapport entre le nombre de motifs corrects retrouvés et le nombre N de réponses du système. Le rappel est la métrique complémentaire qui est le rapport entre le nombre de motifs corrects retrouvés par le système et le nombre total de motifs à retrouver (ici 20). Un motif retrouvé est considéré comme correct si il est distant de moins de 4 trames (ici 64 millisecondes) de la vérité terrain [4]. Plus les réponses d'un système sont pertinentes, plus la précision et le rappel sont élevés.

De la même façon que les courbes ROC, une courbe précision-rappel peut être générée pour une requête et un système en faisant varier un paramètre du système (ici le nombre N de motifs que le système répond). On obtient pour chaque valeur de ce paramètre un couple précision-rappel qui constitue un des points de la courbe précision-rappel. Chaque courbe est composée de 100 points, N variant de 1 à 100. Ainsi, le point le plus à gauche de la courbe est obtenu en ne considérant que la première réponse du système et le point le plus à droite de la courbe en considérant les 100 premières réponses du système. En réalité, les courbes présentées ci-après sont composées des moyennes (pour 50 requêtes) de ces couples précision-rappel. La distance de comparaison choisie et les niveaux des mélanges sont quant eux fixés pour chaque courbe.

La comparaison entre courbes précision-rappel peut s'effectuer de façon globale en comparant les aires sous celles-ci ou encore en regardant quelle est la courbe qui se trouve au dessus des autres sur la totalité du graphe. Cependant, il n'est pas toujours possible de résumer ainsi cette lecture, par exemple lorsque les courbes se croisent. De plus, un algorithme de détection de motifs a plusieurs point de fonctionnement et la comparaison de deux systèmes peut se faire sur la capacité à atteindre de hautes précisions ou de hauts rappels ou sur sa polyvalence.

Limitation de la norme euclidienne l_2 La Figure 5.2 montre les résultats de détection pour la distance l_2 à partir de requête mélangés à un niveau de -12 dB. Les différentes courbes précision-rappel montrent une forte diminution des performances de détection lorsque le niveau des motifs dans les mélanges baisse. Un tel système devient inutilisable pour des valeurs inférieures à 0 dB. C'est pourquoi d'autres distances

4. Lorsqu'un motif correct est détecté, son voisinage (4 trames) est retiré de la liste des segments possibles.

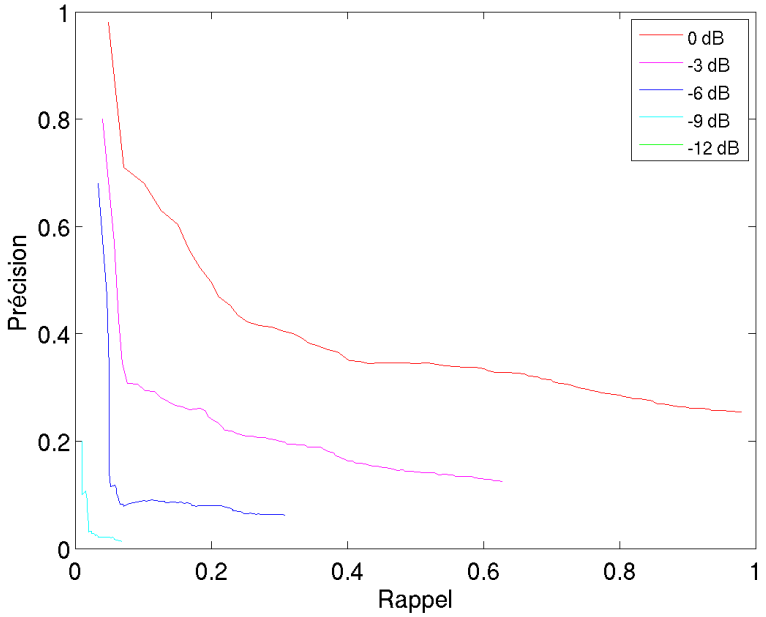


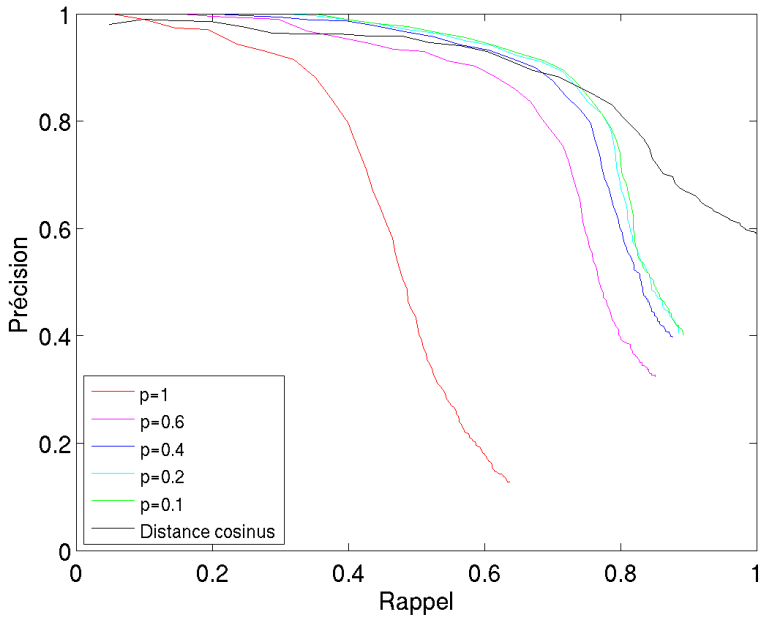
Figure 5.2 – Courbes précision-rappel de la distance l_2 pour différents niveaux (de 0 à -12 dB) des motifs dans les mélanges de recherche pour la tâche de détection de motifs musicaux en présence de parole (niveau des motifs dans la requête : -12 dB).

sont envisagées. A titre de comparaison la courbe « -9 dB » de couleur cyan peut être comparée aux courbes présentes dans la Figure 5.3a.

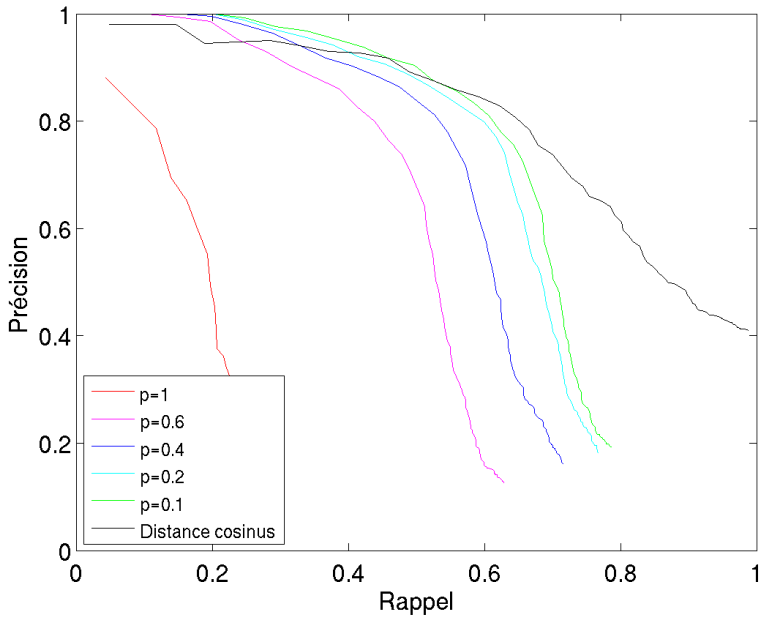
Comparaison des valeurs de p Les valeurs optimales de p établies empiriquement dans la partie précédente étaient de l'ordre de 0,18 pour ce type de donnée quel que soit le niveau considéré. On retrouve cette tendance dans la Figure 5.3 où les courbes des valeurs $p = 0{,}1$ et $p = 0{,}2$ sont supérieures aux autres avec un léger avantage pour $p = 0{,}1$. Plus la valeur de p est faible, plus la détection est performante.

Comparaison à la distance cosinus Lorsque le niveau de la requête et des motifs recherchés est faible (par exemple -12 dB), la distance cosinus permet de maintenir une précision plus haute que les distances l_p lorsque le rappel augmente (voir Figure 5.3). En d'autres termes, la distance cosinus est plus adaptée à des tâches de détection où retrouver l'ensemble des motifs est important.

À l'inverse, les distances l_p atteignent des précisions de 1 ou proches de 1 pour des valeurs faibles de rappel ($< 0{,}4$) contrairement à la distance cosinus (voir Figure 5.3). Ainsi, les distances l_p étudiées offrent une plus grande fiabilité sur les premiers motifs retrouvés (les probabilités de l'hypothèse non-nulle pour le cas à -12 dB sont donnés en Annexe D.2). Cette propriété est par exemple importante pour des tâches de détection qui ne recherchent que quelques motifs avec pour but de les réutiliser pour une autre



(a) Niveau des motifs dans les mélanges de recherche : −9 dB



(b) Niveau des motifs dans les mélanges de recherche : −12 dB

Figure 5.3 – Courbes précision-rappel pour la tâche de détection de motifs musicaux dans de la parole (niveau des motifs dans la requête : −12 dB).

tâche sans avoir besoin de vérifier manuellement leur validité. Des figures complémentaires sont donnés en Annexe B et décrivent le comportement des distances cosinus et $l_{0,1}$ pour différents niveaux dans les mélanges de recherche.

5.3 Conclusion

Nous avons étudié dans ce chapitre les distances l_p pour la comparaison de représentations STFT avec comme application la détection de motifs non déformés en présence d'autres sources. Différents types de motifs et de sources ont été étudiés. Les distances l_p avec $p \leq 0{,}2$ apportent notamment plus de fiabilité aux premiers motifs retrouvés que la distance cosinus au cours de la tâche de détection de motifs musicaux dans de la voix.

5.3.1 Perspectives d'amélioration des distances

Il existe de nombreuses pistes d'amélioration de ces distances, notamment leur utilisation sur les représentations STFT en échelle log ou l'utilisation d'une valeur de p pour les valeurs positives des différences entre STFT et d'une seconde valeur de p pour les valeurs négatives.

Une autre piste serait de normaliser les distances l_p de la même façon que la distance cosinus le fait pour la norme l_2. Il n'existe cependant pas d'équivalent au produit scalaire pour les normes non euclidiennes. Une idée serait l'utilisation de la « distance » suivante

$$d(\mathsf{x},\mathsf{y}) = 1 - \frac{\left(\sum_{f=1}^{F} \text{signe}(\mathsf{x}_f \mathsf{y}_f) |\mathsf{x}_f|^{\frac{p}{2}} |\mathsf{y}_f|^{\frac{p}{2}} \right)^{\frac{2}{p}}}{\|\mathsf{x}\|_p \times \|\mathsf{y}\|_p}, \tag{5.3}$$

qui reste à tester. On retrouve bien la distance cosinus pour $p = 2$.

Enfin, les déformations entre les différentes occurrences d'un motif sont généralement prises en compte par des *features* invariants à ces déformations. Cependant les représentations STFT ne présentant pas souvent cette caractéristique. L'utilisation de ces distances au sein d'une DTW serait un point de départ pour prendre en compte des déformations temporelles. De même considérer des distances l_p non plus pour comparer des trames une à une, mais comparer toutes les trames en même temps favoriserait une certaine parcimonie aussi sur l'axe temporel, mais sans déformations de celui-ci.

5.3.2 Perspectives d'utilisation des distances

Scénario SPORES Dans le cadre de l'approche SPORES et du traitement de bandes-son de films, ces distances permettent de retrouver des motifs identiques en présence d'autres sources dominantes (12 dB). Les requêtes sont en particulier des portions à séparer contenant plusieurs sources, et les motifs identiques retrouvés sont typiquement exploités comme références par une approche de séparation de signal commun (voir chapitre 7) plutôt qu'une approche modélisant les déformations (voir chapitre 6). En

effet, la technique de détection développée dans ce chapitre n'est pas supposée détecter des motifs déformés. De plus, le gain en fiabilité par rapport à la distance cosinus permet un gain de temps de vérification de la validité des motifs retrouvés avant leur utilisation en tant que référence pour la séparation.

Découverte de motifs La tâche de découverte de motifs peut également bénéficier de ces distances afin traiter le cas des mélanges, c'est-à-dire des motifs qui se chevauchent. Le formalisme de découverte de motifs [88] doit cependant être adapté à la présence de plusieurs motifs au même instant.

Application à *REPET-SIM* [145] La propriété de fiabilité sur les premières réponses des distances l_p rendent ces distances intéressantes dans le cas de *REPET-SIM*. En effet, dans cette méthode d'extraction du fond musical, on essaye de détecter pour une trame donnée les trames les plus similaires en présence d'une source variable (la voix chantée). Les distances l_p pourraient alors remplacer la distance cosinus utilisée dans l'approche *REPET-SIM*, si cette propriété de fiabilité se confirme également pour une expérience de détection de motifs de fonds musicaux en présence de voix chantée. L'apprentissage préliminaire de p sur ce type de données (voir Tableau 5.2b) ne laisse pas présager un comportement différent par rapport à la détection de musique dans de la voix (expérience de la partie 5.2). Cette expérience nécessite également l'incorporation d'autres exemples de musique dans la base de recherche, ce qui n'est pas le cas pour le moment et qui représenterait mieux la tâche de détection.

Chapitre 6

Modèle général de déformation pour signaux de référence

Dans ce chapitre, je propose un modèle général de déformation pour la séparation de source guidée par signal de référence déformé. Ce modèle permet l'utilisation conjointe de références multiples, déformées et multicanales. Il permet de modéliser la plupart des problèmes de séparation de sources guidée par signal de référence.

Dans le cas mono-canal, le modèle est estimé par un algorithme de l'état de l'art [65] que j'ai adapté à ce modèle. Je propose également un algorithme GEM inspiré de [136] pour traiter le cas multicanal (voir partie 6.2). Ce modèle et ces algorithmes sont ensuite testés dans différentes configurations

— du modèle de déformation,
— de la phase d'initialisation,
— et du nombre de références,

et sur des tâches

— de modélisation du *pitch-shifting* des références (voir partie 6.3),
— de séparation de mélanges voix/musique guidée par référence (voir partie 6.4),
— et de séparation de morceaux de musique guidée par reprises multi-pistes (voir partie 6.5).

6.1 Modèle général de déformation pour références multiples

Ce modèle comprend plusieurs extensions des formulations et modèles déjà présentés dans l'état de l'art. Ces modifications permettent d'exprimer les liens entre les sources d'un mélange et les signaux de référence et sont :

— une reformulation du problème de séparation à M mélanges,
— le partage des paramètres entre modèles NMF et
— l'ajout de matrices de transformation au modèle NMF excitation-filtre.

6.1.1 Cadre de séparation à M mélanges

Afin de prendre en compte les références, le problème de séparation d'un mélange est reformulé comme le problème de séparation de M mélanges. Ainsi, les signaux contenant les sources de référence sont eux aussi considérés comme des mélanges. Cela permet par exemple de modéliser un signal de référence bruité et d'isoler la source de référence, même si *in fine* on ne sera certainement intéressé que par la séparation d'un mélange en particulier.

Les observations sont donc M mélanges audio $\mathbf{x}^m(t)$ indexés par m. Chaque mélange $\mathbf{x}^m(t)$ est multicanal et contient I^m canaux.

De la même façon que les formulations de l'état de l'art (3.1) et (3.6), chaque mélange est supposé être la somme des images spatiales $\mathbf{y}_j(t)$ d'une ou plusieurs sources indexées par $j \in \mathcal{J}_m$:

$$\mathbf{x}^m(t) = \sum_{j \in \mathcal{J}_m} \mathbf{y}_j(t) \text{ avec } \mathbf{x}^m(t),\mathbf{y}_j(t) \in \mathbb{R}^{I^m}, \tag{6.1}$$

ce qui s'écrit dans le domaine temps-fréquence

$$\mathbf{x}^m_{fn} = \sum_{j \in \mathcal{J}_m} \mathbf{y}_{j,fn} \text{ avec } \mathbf{x}^m_{fn},\mathbf{y}_{j,fn} \in \mathbb{C}^{I^m}, \tag{6.2}$$

avec $f = 1,...,F$ et $n = 1,...,N$ les indices de fréquence et de trame de la STFT. On considère que $\mathbf{x}^1(t)$ est le mélange à séparer et que les mélanges $\mathbf{x}^m(t)$ avec $m > 1$ contiennent les signaux de référence utilisés pour guider la séparation.

La même hypothèse de distribution gaussienne centrée que (3.7) et (3.8) est faite sur les coefficients de la STFT des images spatiales des sources $\mathbf{y}_{j,fn}$:

$$\mathbf{y}_{j,fn} \sim \mathcal{N}_{\mathbb{C}}(0,v_{j,fn}\mathbf{R}_{j,f}) \tag{6.3}$$

avec $v_{j,fn} \in \mathbb{R}_+$ un terme scalaire de puissance spectrale et $\mathbf{R}_{j,f} \in \mathbb{C}^{I^m \times I^m}$ une matrice de covariance spatiale.

Paramètres spatiaux Les matrices de covariance spatiale modélisent les caractéristiques spatiales des sources comme les différences de phase (3.11) et d'intensité (3.12) entre les canaux. Les sources traitées dans les expériences étant spatialement stables au cours du temps, ces matrices seront fixées comme invariantes en temps.

$\mathbf{R}_{j,f}$ est alors représenté comme (3.10) [136], c'est-à-dire $\mathbf{R}_{j,f} = \mathbf{A}_{j,f}\mathbf{A}^H_{j,f}$ avec $\mathbf{A}_{j,f} \in \mathbb{C}^{I^m \times R_j}$ où R_j est le rang de $\mathbf{R}_{j,f}$ et $\mathbf{A}_{j,f}$. La modélisation spectrale étant le point central de ce chapitre, ces aspects ne sont pas plus développés ici.

Paramètres spectraux Le spectrogramme de puissance de chaque source j est noté $V_j = [v_{j,fn}]_{fn} \in \mathbb{R}_+^{F \times N}$ et est modélisé par le modèle excitation-filtre (3.36) :

$$V_j = V_j^e \odot V_j^\phi = W_j^e H_j^e \odot W_j^\phi H_j^\phi. \tag{6.4}$$

Pour rappel, les matrices $W \in \mathbb{R}_+^{F \times D}$ sont des dictionnaires de composantes spectrales, les matrices $H \in \mathbb{R}_+^{D \times N}$ sont les activations temporelles correspondantes, l'exposant ^{e}

est associé à l'excitation, l'exposant $^{\phi}$ est associé au filtre et D^e et D^ϕ sont les nombres de composantes des décompositions NMF.

6.1.2 Modélisation des références déformées

Le modèle général de déformation proposé impose aux matrices W_j^e, H_j^e, W_j^ϕ et H_j^ϕ de suivre un des **trois statuts** suivants indiqués par un code couleur :

- `fixe`, c'est-à-dire inchangée durant l'estimation,
- `libre`, c'est-à-dire adaptée au mélange correspondant m ($j \in \mathcal{J}^m$) durant l'estimation,
- `partagée` [1], c'est-à-dire estimée conjointement entre une source $j \in \mathcal{J}^m$ et une ou plusieurs sources de référence $j' \in \mathcal{J}^{m'}$ avec $m' \neq m$.

Dans ce dernier cas, les déformations entre les sources j et j' sont modélisées par des matrices de transformation $T_{jj'}$. Selon la déformation effective entre la source cible et la source référence, le modèle propose de modéliser ce partage des propriétés spectrales et temporelles entre une source V_j et ses références $V_{j'}$ par une des **trois configurations** illustrées par la Figure 6.1.

Matrices de transformation de l'excitation Concernant la partie excitation, les matrices de transformation sont notées $T_{jj'}^{fe} \in \mathbb{R}_+^{F' \times F}$, $T_{jj'}^{de} \in \mathbb{R}_+^{D^e \times D^e}$ et $T_{jj'}^{te} \in \mathbb{R}_+^{N \times N'}$. Selon la configuration choisie, le partage des composantes spectrales (6.5), des activations temporelles (6.6), ou des deux (6.7) est possible. Ce qui se modélise par une des trois équations suivantes :

$$V_{j'}^e = T_{jj'}^{fe} W_j^e H_{j'}^e \tag{6.5}$$

$$V_{j'}^e = W_{j'}^e H_j^e T_{jj'}^{te} \tag{6.6}$$

$$V_{j'}^e = T_{jj'}^{fe} W_j^e T_{jj'}^{de} H_j^e T_{jj'}^{te}. \tag{6.7}$$

Les matrices de transformation peuvent être `fixes` ou `libres` durant l'étape d'estimation. En pratique, les déformations fréquentielles de l'excitation $T_{jj'}^{fe}$ peuvent par exemple modéliser des différences de vitesse de lecture sur des appareils analogiques ou le changement de dimension spectrale dû à une différence de fréquence d'échantillonnage. $T_{jj'}^{te}$ est utilisée pour aligner temporellement les spectres des signaux et représente le chemin d'alignement temporel entre deux signaux. $T_{jj'}^{de}$ modélise les changements internes au dictionnaire d'excitation, comme par exemple le *pitch shifting* [2]. Cette dernière matrice de transformation n'apparait que quand les W_j,H_j correspondantes sont `partagées`, dans le cas inverse le modèle est redondant.

1. Lorsque plusieurs couples de matrices sont présents dans la même équation, un des couples est indiqué en bleu (`partagée`).

2. On peut noter que le *pitch shifting* et la différence de vitesse de lecture ont deux effets différents, en particulier pour les sons inharmoniques.

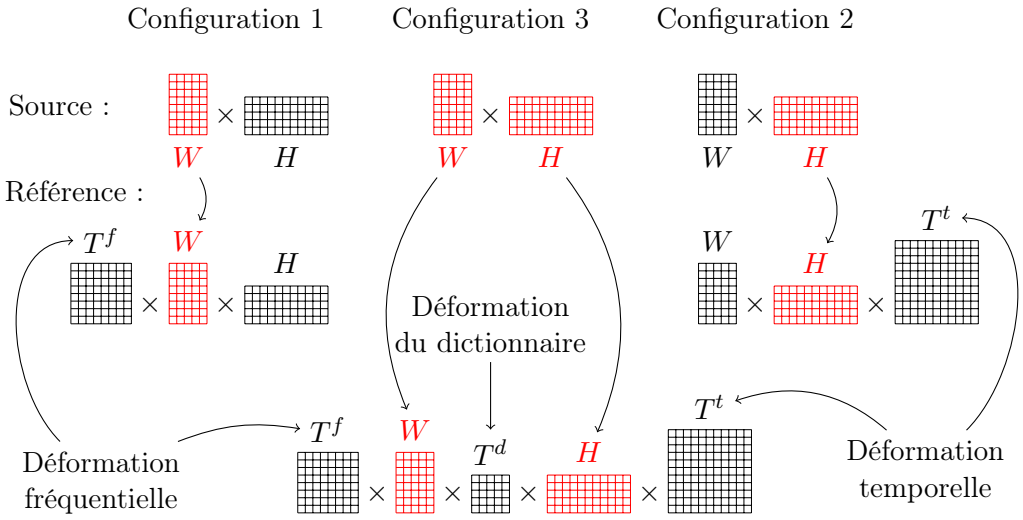


Figure 6.1 – Illustration des trois configurations du modèle général de déformation.

Matrices de transformation du filtre En ce qui concerne le filtre, les matrices de transformation entre la source cible et la source de référence sont notées $T^{f\phi}_{jj'} \in \mathbb{R}_+^{F'\times F}$, $T^{d\phi}_{jj'} \in \mathbb{R}_+^{D^\phi\times D^\phi}$ et $T^{t\phi}_{jj'} \in \mathbb{R}_+^{N\times N'}$. De la même façon que pour la partie excitation, les trois configurations possibles permettent de partager les composantes spectrales (6.8), les activations temporelles (6.9), ou les deux (6.10) :

$$V^\phi_{j'} = T^{f\phi}_{jj'} W^\phi_j H^\phi_{j'} \tag{6.8}$$
$$V^\phi_{j'} = W^\phi_{j'} H^\phi_j T^{t\phi}_{jj'} \tag{6.9}$$
$$V^\phi_{j'} = T^{f\phi}_{jj'} W^\phi_j T^{d\phi}_{jj'} H^\phi_j T^{t\phi}_{jj'}. \tag{6.10}$$

Comme précédemment, les matrices de transformation sont soit `fixes` ou `libres`. Les déformations fréquentielles du filtre $T^{f\phi}_{jj'}$ peuvent par exemple modéliser les changements de longueur de conduit vocal [106] ou des différences d'égalisation. $T^{d\phi}_{jj'}$ modélise les changements internes au dictionnaire du filtre, comme par exemple le changement d'un phonème lorsque le locuteur a un accent différent, et qu'un phonème est toujours prononcé à la place d'un autre. De même que pour l'excitation, cette dernière matrice n'apparait que lorsque les W et H correspondantes sont `partagées`. $T^{t\phi}_{jj'}$ modélise les déformations temporelles du filtre, et est utilisée pour aligner temporellement les signaux.

La Figure 6.2 donne une illustration d'une possible utilisation de ce modèle. Elle correspond à une référence de parole modélisée par (6.26) et une source cible de parole qui a été produite par un locuteur différent et qui est présente dans un mélange (6.25). Plus de détails sur cet exemple sont donnés dans la partie 6.4.2.

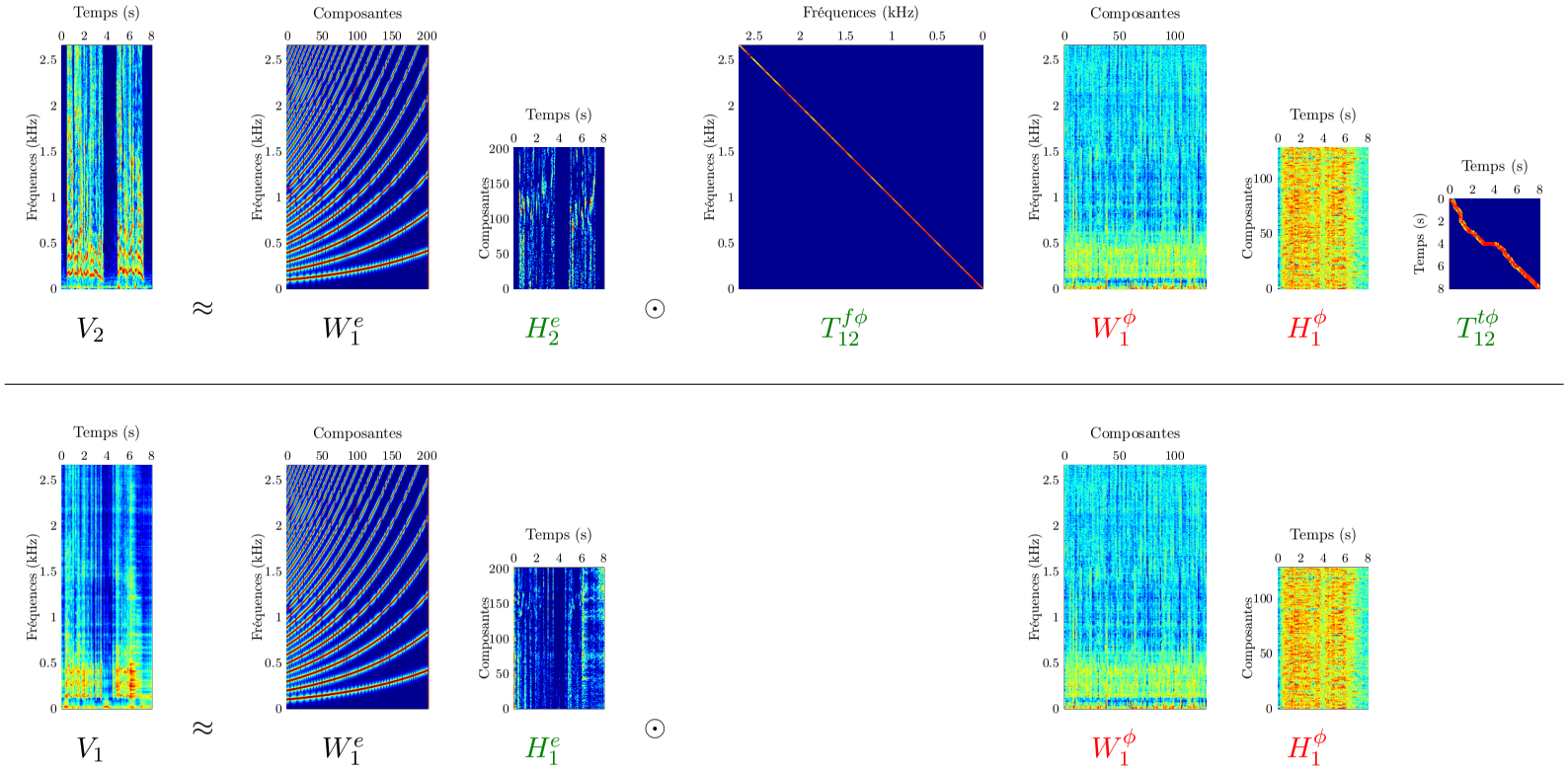


Figure 6.2 – Exemple de décomposition du spectre de puissance d'un mélange de référence ($m' = 2$) contenant une seule source ($j' = 2$) similaire à la source de parole $j = 1 \in \mathcal{J}^1$. Les paramètres en noir sont `fixes`, ceux en vert sont `libres` et ceux en rouge sont `partagés`. Les paramètres de la source cible ($j = 1$) sont également affichés.

6.1.3 Discussion

Comparaison avec d'autres approches Ce modèle général de déformation généralise les approches de l'état de l'art [106, 154, 160] qui exploitent des modèles similaires. Dans [154], la source de référence est composée de notes de musique isolées et peut être modélisée sans modèle source-filtre par le `partage` des matrices W_j et des matrices H_j `libres`. L'approche décrite dans [106] est exactement exprimée par (6.25) et (6.26).

Le modèle proposé peut aussi représenter les propriétés des signaux utilisées dans d'autres approches [50, 75, 85, 106, 113, 154, 157, 162] bien que les modèles soient quelque peu différents. L'approche de séparation par fredonnement [157] peut être implémentée par le `partage` de la partie excitation et par une partie filtre `libre`. Les informations symboliques de musique (resp. de parole) peuvent être utilisées après avoir été synthétisées comme dans [71, 74] (resp. [106]), ou directement dans le modèle [50, 154] en contraignant H_j^e (resp. H_j^ϕ).

Enfin, ce modèle ouvre la voie à de nouveaux scénarios comme par exemple la séparation d'un couplet de morceau de musique guidée par un autre couplet. Dans ce cas, la source de voix chantée aurait des matrices d'excitation (H_j^e and W_j^e) `partagées` mais un filtre différent (H_j^ϕ) au cours du temps. L'approche serait alors similaire à REPET [146] mais la voix serait considérée comme un motif répété, alors que REPET modélise uniquement un fond musical sans déformation.

Extensions de l'approche Comme déjà mentionné, ce modèle suppose que la cible est indexée par $j \in \mathcal{J}^1$ et la référence par $j' \in \mathcal{J}^{m'}$ avec $m' \neq 1$ ce qui est censé représenter le scénario classique de la séparation guidée par référence. Retirer cette contrainte mène à de nouvelles possibilités pour la séparation.

Par exemple modéliser la relation entre des sources du même mélange, c'est-à-dire j,$j' \in \mathcal{J}^m$ peut avoir un intérêt en présence d'un délai entre sources du même mélange comme dans le cas d'un canon en musique. La prise en compte de relations circulaires, c'est-à-dire $T_{jj'}$,$T_{j'j''}$,$T_{j''j}$ permettrait la séparation conjointe des différents mélanges, mais introduirait une matrice de transformation supplémentaire. Plus généralement, considérer le mélange à séparer comme central est une bonne stratégie pour éviter l'ajout de matrices et de potentiel effet de lissage sur les sources cibles.

Une autre extension possible serait d'autoriser le partage des matrices de transformation, ce qui pourrait être utile lorsque plusieurs instruments subissent la même transformation (tonalité ou *pitch shifting*) ou lorsque l'excitation et le filtre sont assujettis à la même déformation temporelle [3].

Contraintes supplémentaires Lorsque l'on utilise un modèle excitation-filtre, l'estimation du filtre requiert généralement l'ajout de contraintes de régularité (*smoothness*), comme dans [48, 136], afin de garantir que le filtre représente bien la résonance du conduit vocal ou de l'instrument de musique. Dans les expériences présentées dans

3. Ces matrices de transformation `partagées` serait alors estimées en suivant (6.15) ou (6.17) de la même façon que les autres paramètres spectraux `partagés`.

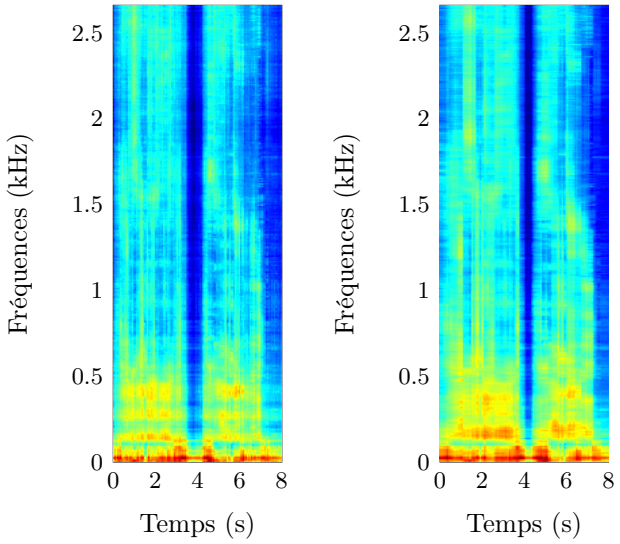


Figure 6.3 – Exemples de filtres V^{ϕ} estimés pour une source de parole (6.3a), et sa référence (6.3b) dans les expériences de la partie 6.4.

ce chapitre, aucune contrainte de ce type n'est utilisée pour ne pas surcharger le modèle. J'ai cependant pu observer expérimentalement que les filtres étaient réguliers lorsqu'ils sont estimés en présence de contraintes provenant de références. La Figure 6.3 montre des exemples de filtres obtenus au cours des expériences décrites dans la partie 6.4. Une possible explication est qu'utiliser plus qu'une source pour l'estimation d'un filtre donné (alors que l'excitation est différente pour chaque source) procure plus de robustesse à l'estimation plutôt que de n'utiliser qu'une seule source. Il est toutefois difficile, sinon impossible, d'apporter des garanties théoriques de ce comportement.

Si nécessaire, des contraintes explicites de régularité pourraient être incluses dans le modèle général précédemment présenté soit en contraignant les matrices W_j^{ϕ} comme devant être le produit de composantes fréquentielles locales régulières et de coefficients d'enveloppe spectrale [48, 136] soit en introduisant des à priori probabilistes sur les coefficients de W_j^{ϕ} [129]. De la même façon, des contraintes de continuité temporelle peuvent être imposées sur H_j^{ϕ} ou H_j^{e}. Ces contraintes ne sont cependant pas étudiées dans ce travail.

6.2 Estimation des paramètres

Je présente ci-après deux méthodes d'estimation des paramètres au sens du maximum de vraisemblance, ce qui peut s'écrire comme :

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda^m \log p(\mathbf{x}^m|\theta) \tag{6.11}$$

où θ est l'ensemble des paramètres à estimer, c'est-à-dire la matrice de mélange $\mathbf{A}_f$, et les matrices W, H et T qui sont soit `libres` soit `partagées`. Les $\lambda^m \in \mathbb{R}_+$ pondèrent les potentielles différences de durée ou de résolution fréquentielle entre les mélanges 1 et m, ou permettent de donner plus d'importance aux références supposées les plus importantes. L'influence de ces λ a été étudiée dans d'autres travaux [105].

Je présente tout d'abord un algorithme basé sur des mises à jour multiplicatives (*Multiplicative Updates* ou MU) pour traiter le cas mono-canal. Ensuite, un algorithme GEM est proposé pour l'estimation des paramètres dans le cas multicanal. Enfin, différentes procédures d'initialisation sont présentées.

6.2.1 Mises à jour multiplicatives dans le cas mono-canal

Dans le cas mono-canal, maximiser la log-vraisemblance est équivalent [65] à minimiser la divergence d'IS (3.21) :

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{m=1}^{M} \lambda^m \sum_{f,n=1}^{F,N} d_{IS}(X_{fn}^m | V_{fn}^m) \tag{6.12}$$

où $X^m = [|\mathbf{x}_{fn}^m|^2]_{fn}$ et $V^m = \sum_{j \in \mathcal{J}^m} V_j$ sont respectivement les spectrogrammes de puissance observés et estimés. D'autres divergences sont envisageables cependant elles ne sont pas extensibles au cas multicanal contrairement à IS.

Des mises à jour multiplicatives (3.24) sont appliquées itérativement à chaque paramètre afin de faire diminuer le critère (6.12) [66]. Selon le statut (`libre` ou `partagé`), différentes mises à jour multiplicatives sont obtenues pour chaque paramètre.

Pour les paramètres `libres` (en vert), cela conduit à la mise à jour classique de la NMF. Un exemple d'une telle mise à jour est donné par (6.14) pour le paramètre W_j^e. Pour les paramètres `partagés` (en rouge), cela conduit aux mises à jour de la NMPcF. Un exemple d'une telle mise à jour est donné par (6.15) [4] pour le paramètre W_j^e.

6.2.2 Algorithme GEM pour le cas multicanal

Dans le cas multicanal, la séparation de sources guidée par référence peut aussi exploiter l'information spatiale apportée par les différents canaux, en particulier si les sources ont des angles d'arrivée différents. L'intérêt des données multicanales reste valable même si les mélanges ont des nombres de canaux différents ou si aucune hypothèse de similarité entre les directions d'arrivée des sources cibles et de leur références n'est faite.

En suivant le cadre de modélisation des mélanges multicanaux proposé dans [136] et rappelé dans la partie 3.3.2.2, la prise en compte de M mélanges implique que (3.26) s'exprime pour chaque mélange :

$$\mathbf{x}_{fn}^m = \mathbf{A}_f^m \mathbf{s}_{fn}^m + \mathbf{b}_{fn}^m \tag{6.13}$$

4. Dans (6.15) et (6.17), nous avons supposé que $V_{j'}^e$ et $V_{j'}^\phi$ suivent les modèles (6.7) et (6.10). En pratique, le nombre de paramètres `partagés` et de matrices de transformation est généralement plus faible, comme illustré dans les parties 6.4 et 6.5.

MU - paramètre `libre` :

$$W_j^e \leftarrow W_j^e \odot \frac{[V_j^\phi \odot V^{m\cdot[-2]} \odot X^m][H_j^e]^T}{[V_j^\phi \odot V^{m\cdot[-1]}][H_j^e]^T} \tag{6.14}$$

MU - paramètre `partagé` :

$$W_j^e \leftarrow W_j^e \odot \frac{\lambda^m[V_j^\phi \odot V^{m\cdot[-2]} \odot X^m][H_j^e]^T + \sum_{j'} \lambda^{m'}[T_{jj'}^{fe}]^T[V_{j'}^\phi \odot V^{m'\cdot[-2]} \odot X^{m'}][T_{jj'}^{de} H_j^e T_{jj'}^{te}]^T}{\lambda^m[V_j^\phi \odot V^{m\cdot[-1]}][H_j^e]^T + \sum_{j'} \lambda^{m'}[T_{jj'}^{fe}]^T[V_{j'}^\phi \odot V^{m'\cdot[-1]}][T_{jj'}^{de} H_j^e T_{jj'}^{te}]^T} \tag{6.15}$$

EM - paramètre `libre` :

$$W_j^e \leftarrow W_j^e \odot \frac{[V_j^\phi \odot V_j^{\cdot[-2]} \odot \hat{\Xi}_j][H_j^e]^T}{[V_j^\phi \odot V_j^{\cdot[-1]}][H_j^e]^T} \tag{6.16}$$

EM - paramètre `partagé` :

$$W_j^e \leftarrow W_j^e \odot \frac{\lambda^m R_j[V_j^\phi \odot V_j^{\cdot[-2]} \odot \hat{\Xi}_j][H_j^e]^T + \sum_{j'} \lambda^{m'} R_{j'}[T_{jj'}^{fe}]^T[V_{j'}^\phi \odot V_{j'}^{\cdot[-2]} \odot \hat{\Xi}_{j'}][T_{jj'}^{de} H_j^e T_{jj'}^{te}]^T}{\lambda^m R_j[V_j^\phi \odot V_j^{\cdot[-1]}][H_j^e]^T + \sum_{j'} \lambda^{m'} R_{j'}[T_{jj'}^{fe}]^T[V_{j'}^\phi \odot V_{j'}^{\cdot[-1]}][T_{jj'}^{de} H_j^e T_{jj'}^{te}]^T} \tag{6.17}$$

où $\mathbf{b}_{fn}^m$ est un bruit additif isotrope de covariance diagonale $\Sigma_{\mathbf{b}_{fn}^m} = \sigma_f^2 \mathbf{I}_{I^m} \in \mathbb{C}^{I^m \times I^m}$ et $\mathbf{A}_f^m \in \mathbb{C}^{I^m \times R^m}$ (resp. $\mathbf{s}_{fn}^m \in \mathbb{C}^{R^m}$) [5] résulte de la concaténation des matrices de mélange $\mathbf{A}_{j,f}$ (resp. de toutes les sous-sources $s_{jr,fn}$) de toutes les sources $j \in \mathcal{J}^m$. La log-vraisemblance des données complètes s'écrit maintenant :

$$Q(\theta,\theta^c) \stackrel{c}{=} -\sum_{mfn} \frac{\lambda^m}{\sigma_f^2} \mathrm{tr}\left[\widehat{\mathbf{R}}_{\mathbf{x}_{fn}^m} - \mathbf{A}_f^m \widehat{\mathbf{R}}_{\mathbf{x}\mathbf{s}_{fn}^m}^H - \widehat{\mathbf{R}}_{\mathbf{x}\mathbf{s}_{fn}^m} \mathbf{A}_f^{mH} + \mathbf{A}_f^m \widehat{\mathbf{R}}_{\mathbf{s}_{fn}^m} \mathbf{A}_f^{mH}\right] - \sum_{m,j\in\mathcal{J}^m,fn} \lambda^m R_j d_{IS}(\widehat{\xi}_{j,fn}|v_{j,fn}), \quad (6.18)$$

avec : $\widehat{\mathbf{R}}_{\mathbf{x}_{fn}^m} \triangleq \widehat{\mathbb{E}}[\mathbf{x}_{fn}^m \mathbf{x}_{fn}^{m\,H}]$, $\widehat{\mathbf{R}}_{\mathbf{x}\mathbf{s}_{fn}^m} \triangleq \widehat{\mathbb{E}}[\mathbf{x}_{fn}^m \mathbf{s}_{fn}^{m\,H}|\theta^c]$, $\widehat{\mathbf{R}}_{\mathbf{s}_{fn}^m} \triangleq \widehat{\mathbb{E}}[\mathbf{s}_{fn}^m \mathbf{s}_{fn}^{m\,H}|\theta^c]$ et $\widehat{\xi}_{j,fn} \triangleq \frac{1}{R_j}\sum_{r=1}^{R_j} \widehat{\mathbb{E}}\left[|s_{jr,fn}|^2|\theta^c\right]$. L'obtention de la log-vraisemblance (6.18) est détaillée en Annexe C.1.

L'algorithme GEM alterne ensuite les étapes « E-step » et « M-step » suivantes afin de faire croitre la vraisemblance Q.

E-step Cette étape consiste à calculer les statistiques suffisantes $\widehat{\mathbf{R}}_{\mathbf{s}_{fn}^m} \in \mathbb{C}^{R^m \times R^m}$ et $\widehat{\mathbf{R}}_{\mathbf{x}\mathbf{s}_{fn}^m} \in \mathbb{C}^{I^m \times R^m}$ de la même façon que (3.28) et (3.29) pour chaque mélange.

M-step Si aucun des paramètres n'est `partagé`, l'algorithme GEM traite les différents mélanges séparément et se comporte comme dans [136].

Les paramètres spatiaux `libres` présents dans l'ensemble θ sont mis à jour afin de faire croitre le premier terme de (6.18) en annulant sa dérivée par rapport à $\mathbf{A}_f$ (3.33) [136] :

$$\mathbf{A}_f = \left[\sum_n \widehat{\mathbf{R}}_{\mathbf{x}\mathbf{s}_{fn}^m}\right] \left[\sum_n \widehat{\mathbf{R}}_{\mathbf{s}_{fn}^m}\right]^{-1}. \quad (6.19)$$

Les paramètres spectraux sont quant à eux mis à jour afin de minimiser

$$\sum_{m,j\in\mathcal{J}^m,fn} R_j d_{IS}(\widehat{\xi}_{j,fn}|v_{j,fn}) = \sum_{m,j\in\mathcal{J}^m} R_j D_{IS}(\widehat{\Xi}_j|V_j) \quad (6.20)$$

avec $\widehat{\Xi}_j = [\widehat{\xi}_{j,fn}]_{fn} \in \mathbb{R}_+^{F\times N}$ où $\widehat{\xi}_{j,fn} = \frac{1}{R_j}\sum_{r=1}^{R_j} \widehat{\mathbf{R}}_{\mathbf{s}_{fn}^m}(r,r)$.

Le partage des paramètres spectraux induit un simple changement durant l'étape M de ces paramètres spectraux `partagés`. Des exemples de mises à jour multiplicatives sont données pour les paramètres `libres` (en vert) par (6.16) et pour les paramètres `partagés` (en rouge) par (6.17) [6]. Ils généralisent la mise à jour (30) dans [136].

5. $R^m = \sum_{j\in\mathcal{J}^m} R_j$
6. voir [4] page 68

6.2.3 Initialisation des paramètres

Les résultats de ces deux algorithmes dépendent grandement de l'initialisation. Par rapport à la séparation de sources aveugle ou faiblement guidée, la séparation de sources guidée par signal de référence offre la possibilité d'avoir des meilleurs valeurs initiales pour les paramètres W et H en tirant parti des références disponibles. Les autres paramètres, c'est-à-dire les matrices de transformation T, sont au préalable grossièrement estimées (voir partie 6.4). On peut par exemple utiliser des mises à jour multiplicatives pour minimiser le critère suivant :

$$\widehat{\theta}_{\text{ref}} = \underset{\theta_{\text{ref}}}{\operatorname{argmin}} \sum_{m=2}^{M} \lambda^m \sum_{f,n=1}^{F,N} d_{IS}(X_{fn}^m | V_{fn}^m) \tag{6.21}$$

où θ_{ref} est l'ensemble des paramètres W et H qui apparaissent dans les mélanges de référence indexés par m,$2 \geq m \geq M$. Ceci est particulièrement efficace lorsqu'une seule source domine dans le mélange de référence. À la fin de cette étape, les paramètres du mélange principal qui ne sont pas `partagés` seront les seuls à rester mal initialisés (si aucune autre information à priori n'est disponible à leur sujet).

Les étapes algorithmiques suivantes seront utilisées dans les expériences :

- *Init* : les statuts des paramètres sont définis pour chaque source, c'est-à-dire, `fixe`, `libre` ou `partagé` et ils sont initialisés suivant les informations à priori à disposition. Les détails de cette étape sont donnés pour chaque expérience dans les parties 6.4 et 6.5 selon le scénario envisagé,
- *NMF* : les W, H `partagés` ou `libres` des mélanges de référence sont mis à jour par MU, c'est-à-dire comme décrit ci-avant dans cette partie 6.2.3 pour minimiser (6.21) [7],
- *Plain-NMF* : l'algorithme décrit dans la partie 6.2.1 est appliqué au mélange principal uniquement ($M = 1$ dans (6.12)),
- *NMPcF* : l'algorithme décrit dans la partie 6.2.1 est appliqué à tous les mélanges,
- *GEM* : l'algorithme décrit dans la partie 6.2.2 est appliqué à tous les mélanges.

Dans les expériences qui vont suivre, différentes combinaisons de ces quatre étapes algorithmiques seront testées dans l'ordre présenté ci-dessus. Dans tous les cas, l'estimation finale des sources est effectuée par un filtre de Wiener adaptatif (3.15).

6.3 Scénario élémentaire avec *pitch shifting*

Dans cette partie, le modèle général de déformation est appliqué sur des signaux comportant une seule source et pour laquelle une version « *pitch shiftée* » sert de référence. Cet exemple élémentaire permet d'illustrer une utilisation possible des matrices de transformation de la partie excitation T^{fe} et T^{de}.

7. En fait, si un paramètre spectral est `partagé` entre plusieurs mélanges de référence des MU plus similaires à (6.15) sont utilisées.

6.3.1 Données

Les données de test sont composées de six extraits de guitare d'une durée de trente secondes ainsi que les références « *pitch shiftées* » (de un à quatre demi-tons) pour chacun de ces extraits. Les références « *pitch shiftées* » sont générées artificiellement à partir des extraits en utilisant le GuitarPitchShifter [8].

6.3.2 Modèle et initialisation

La source est ci-après numérotée $j = 1$ et la référence $j = 2$. Dans cette description, j'ai retiré la notion de mélange puisque les signaux ne contiennent qu'une seule source. Les variables `fixes` sont en noir (W_1^e), les variables `libres` sont en vert ($T_{12}^{f\phi}$, $T_{12}^{d\phi}$) et les variables `partagée` sont en rouge (H_1^e, W_1^ϕ, H_1^ϕ). Le spectre de puissance de la source V_1 est modélisé comme

$$V_1 = W_1^e H_1^e \odot W_1^\phi H_1^\phi \tag{6.22}$$

et le spectre de puissance de la référence V_2 est modélisé soit par

$$V_2 = W_1^e T_{12}^{de} H_1^e \odot W_1^\phi H_1^\phi. \tag{6.23}$$

soit par

$$V_2 = T_{12}^{fe} W_1^e H_1^e \odot W_1^\phi H_1^\phi \tag{6.24}$$

La matrice des composantes spectrales de l'excitation W_1^e est un dictionnaire `fixe` de composantes harmoniques calculées comme dans [136] (voir aussi la Figure 6.2). Chaque composante est un spectre harmonique et deux composantes successives sont séparées par un demi-ton. De façon à représenter la transformation induite par le *pitch shifting*, deux modèles alternatifs ont été testés :

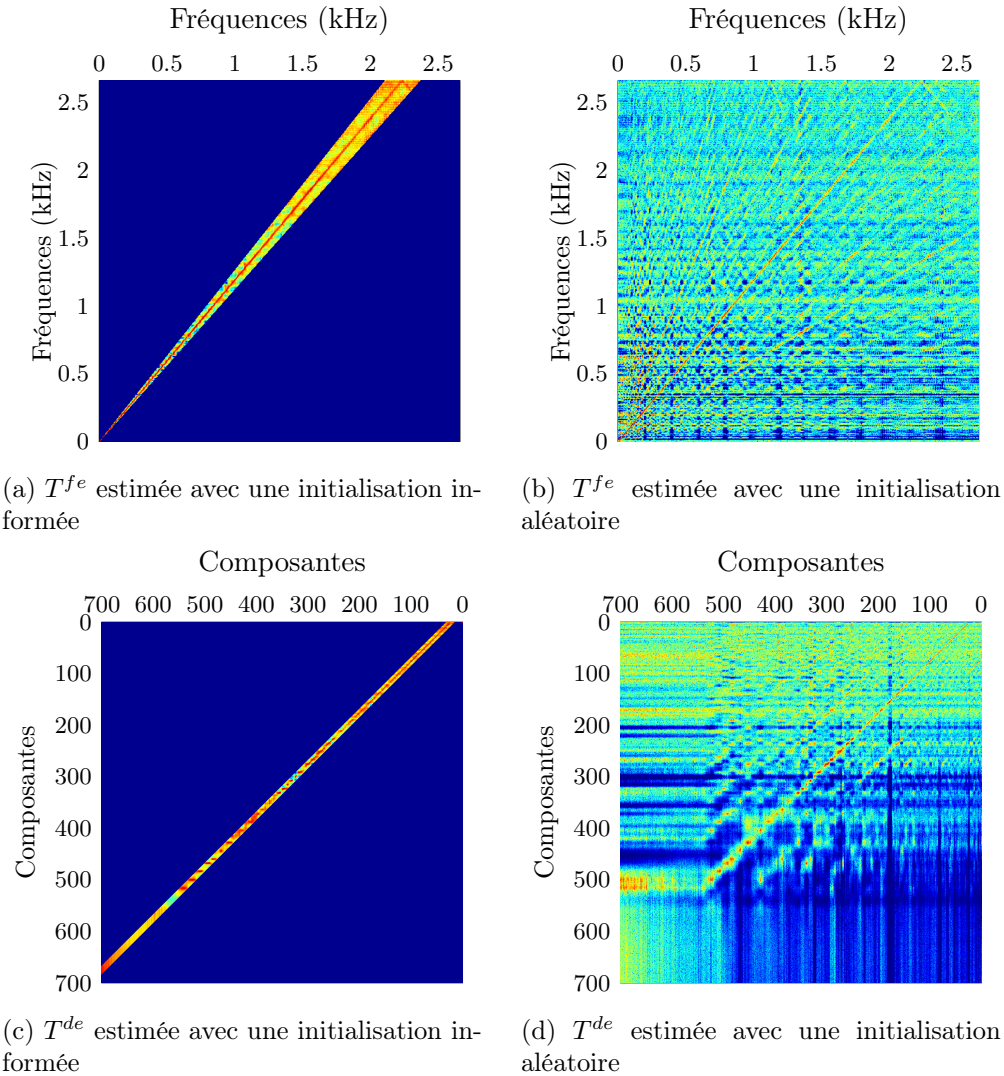
- une transformation du dictionnaire des composantes (T^{de}) (6.23) qui devrait être, dans le cas idéal, une translation définie par l'équation $y = x + b$, où b est la valeur du *pitch shifting* en demi-tons.
- une transformation fréquentielle (T^{fe}) (6.24) qui devrait être, dans le cas idéal, une homothétie définie par l'équation $y = \alpha x$, où $\alpha = 2^{b/12}$.

Ces matrices de transformation `libres` $T_{12}^{f\phi}$ et $T_{12}^{d\phi}$ sont soit initialisées

- de façon informée, c'est-à-dire que les éléments de la matrice à moins d'un ton du véritable *pitch shift* sont initialisés avec des valeurs aléatoires et les autres éléments sont mis à zéro,
- soit entièrement avec des valeurs aléatoires.

Des exemples de matrices de transformation estimées sont illustrés dans la Figure 6.4. On peut rappeler qu'avec les mises à jour multiplicatives les zéros (bleu foncé dans les figures) restent inchangés au fil des itérations. Tous les éléments des autres matrices (H_1^e, W_1^ϕ, H_1^ϕ) sont initialisés aléatoirement.

8. http ://www.guitarpitchshifter.com/matlab.html



(a) T^{fe} estimée avec une initialisation informée

(b) T^{fe} estimée avec une initialisation aléatoire

(c) T^{de} estimée avec une initialisation informée

(d) T^{de} estimée avec une initialisation aléatoire

Figure 6.4 – Exemples de différentes matrices de déformation modélisant le *pitch shifting* d'un extrait de guitare.

***Pitch shift* (tons)**	1/2		1		3/2		2	
Source/Référence	Src	Ref	Src	Ref	Src	Ref	Src	Ref
Oracle	10,3	9,6	10,3	9,6	10,2	9,4	10,2	9,6
T^f **informé**	9,1	7,9	8,8	7,6	8,5	7,5	8,5	7,3
T^f **aléatoire**	9,0	7,8	8,7	7,4	8,3	7,1	8,1	7,0
T^d **informé**	8,5	7,2	7,9	6,9	7,5	6,6	7,4	6,2
T^d **aléatoire**	9,3	6,9	8,7	6,5	8,5	6,2	8,3	5,5
Sans déformation	6,5	4,5	6,2	3,7	6,0	3,1	5,9	2,8

(a) Moyennes des SNR (dB).

***Pitch shift* (tons)**	1/2		1		3/2		2	
Source/Référence	Src	Ref	Src	Ref	Src	Ref	Src	Ref
Oracle	0,09	0,10	0,09	0,10	0,09	0,10	0,09	0,11
T^f **informé**	0,13	0,19	0,15	0,20	0,15	0,21	0,15	0,22
T^f **aléatoire**	0,12	0,19	0,13	0,20	0,13	0,21	0,13	0,22
T^d **informé**	0,13	0,19	0,14	0,20	0,15	0,22	0,16	0,23
T^d **aléatoire**	0,13	0,20	0,14	0,21	0,15	0,23	0,16	0,24
Sans déformation	0,18	0,28	0,20	0,28	0,24	0,28	0,27	0,29

(b) Moyennes des divergences d'IS.

Tableau 6.1 – Rapport signal-à-bruit et divergence d'IS entre les spectres estimés et observés pour le scénario élémentaire avec des extraits de guitare *pitch-shiftés* comme décrit dans la partie 6.3.

6.3.3 Estimation et résultats

L'estimation des paramètres du modèle conjoint, c'est-à-dire des équations (6.22) et (6.23) ou (6.22) et (6.24), est effectuée en utilisant des mises à jour de type *NMPcF* (voir partie 6.2.1). Une expérience sans déformation dans le modèle de la référence, c'est-à-dire $V_2 = V_1$, est ajoutée pour comparaison, de même qu'une expérience « oracle », c'est-à-dire $V_2 = T_{12}^{fe} W_1^e H_2^e \odot W_2^\phi H_2^\phi$ ou bien $V_2 = W_1^e T_{12}^{de} H_2^e \odot W_2^\phi H_2^\phi$. L'expérience oracle correspond au cas où aucun paramètre n'est partagé entre les modèles des spectres de puissance de la source et de la référence. En d'autres termes, les modèles sont estimés séparément pour s'ajuster à l'observation des vraies sources. Ces deux expériences complémentaires sont censées respectivement apporter des bornes inférieures et supérieures de performance à notre approche. Les résultats sont donnés en terme de rapport signal-à-bruit (*Signal-to-Noise Ratio* ou SNR) entre les spectres d'amplitude des observations ($[|\mathbf{x}_{fn}^1|]_{fn}$ et $[|\mathbf{x}_{fn}^2|]_{fn}$) et des spectres estimés ($V_1^{\cdot[1/2]}$ et $V_2^{\cdot[1/2]}$) dans le Tableau 6.1a et en terme de la divergence d'IS (les termes de la somme dans (6.12)) dans le Tableau 6.1b.

Que ce soit pour le SNR ou la divergence d'IS, on observe que la distorsion est bien plus petite après avoir été modélisée, par exemple 9,1 dB et 7,9 dB SNR au lieu 6,5 dB et 4,5 dB (première colonne dans le Tableau 6.1a). Néanmoins, la distorsion reste légèrement plus importante que dans ce qui est obtenu avec l'expérience oracle, par exemple 10,3 dB et 9,6 dB. Ces résultats montrent la capacité des différents modèles proposés à prendre en compte le *pitch shifting* et à effectivement réduire la différence entre les signaux et les modèles correspondants. On peut aussi noter que la connaissance à priori de la valeur du *pitch shifting* permet une légère amélioration en terme de SNR. Une telle information peut par exemple être fournie par un ingénieur du son ou un musicien.

Les *pitch shifting* étudiés ici étant produits artificiellement par un logiciel, l'utilisation de T^{fe} est possible bien que la guitare soit inharmonique. Dans un scénario différent où une mélodie aurait été jouée par un instrument inharmonique à deux tonalités différentes, l'inharmonicité aurait requis un dictionnaire inharmonique spécifique et l'utilisation de T^{de} au lieu de T^{fe}. En effet, les partiels d'un *pitch* donné n'auraient pas été retrouvés par une simple translation des partiels d'un autre *pitch*. Les composantes « pitchées » doivent être translatées alors que les composantes d'attaque ou percussives ne doivent pas être modifiées. Une telle distinction pour chaque composante n'est pas possible avec T^{fe}.

6.4 Séparation voix/musique

Dans cette partie, je décris un second cas d'utilisation du modèle proposé pour la séparation de sources guidée par signal référence déformé. L'objectif est la séparation de la voix et de la musique dans des bandes-son de films et de séries télé ayant été produites de façon analogique. Des références de voix et/ou de musique sont disponibles et permettent de guider la séparation.

Après avoir rapidement décrit les données, j'expose comment les références de voix et de musique sont modélisées dans le cadre du modèle général proposé. Différentes

combinaisons algorithmiques sont évaluées et l'utilisation de plusieurs références pour une même source est étudiée dans le but d'améliorer la qualité de la séparation.

6.4.1 Données

Les extraits musicaux ainsi que les références correspondantes sont obtenus à l'aide du logiciel de découverte de motifs MODIS [34] basé sur les travaux de MUSCARIELLO [123]. Ce logiciel vise à agréger les segments d'un long flux audio (ici des films et des séries télé) qui sont suffisamment similaires étant donné un seuil préétabli (voir partie 2.4.3). Une certaine variabilité étant tolérée par la découverte de motifs, les répétitions découvertes peuvent donc être non exactes, c'est-à-dire déformées par rapport à la source cible (changement de rythme, *fade in*) et contenir des sources supplémentaires (principalement des effets sonores ou des bruitages).

Les exemples de parole et leur références sont quant à eux issus d'une base de données déjà existante [17] dans laquelle 16 locuteurs différents prononcent les 238 mêmes phrases. J'ai conservé 4 exemples musicaux et 4 phrases (deux locuteurs femmes et deux locuteurs hommes) afin de générer les mélanges à deux rapports voix/musique différents : -6 dB (la musique au premier plan et la voix en arrière-plan), et 12 dB (la voix au premier plan et la musique en arrière-plan). Ainsi, les SDR initiaux sont -6 dB et 12 dB pour la voix et 6 dB et -12 dB pour la musique. Ces niveaux sont proches de ceux effectivement observés dans les films et les séries télé. Les mélanges sont ainsi obtenus afin d'évaluer objectivement les résultats par rapport à la vérité terrain à l'aide de [174]. La combinaison de tous ces paramètres permet de produire 32 mélanges originaux X^1. Pour chaque mélange à séparer, on dispose de plusieurs références de musique (les autres extraits découverts), et plusieurs référence de voix (les mêmes phrases prononcées par d'autres locuteurs). Le nombre de références utilisées varie au cours des expériences. Les mélanges originaux et les références ont une durée d'environ 8 secondes. Ils sont mono-canal ($\forall m, I^m = 1$) et échantillonnés à 16 kHz. Des exemples audio sont disponibles en ligne [9].

6.4.2 Modèles testés

Dans les différentes configurations qui vont suivre, les sources de parole sont numérotées $j = 1$ ou 2, les sources de musique $j = 3$ ou 4, et les autres sources et le bruit de fond $j = 5$ et 6. Les variables `fixes` sont en noir (W_1^e,W_3^e,$T_{34}^{t\phi}$). Les variables `libres` sont en vert (H_1^e, H_2^e, $T_{12}^{f\phi}$, $T_{12}^{d\phi}$, $T_{12}^{t\phi}$, T_{34}^{te}, W_5, H_5, W_6, H_6). Les variables `partagées` sont en rouge ou en bleu (W_1^ϕ, H_1^ϕ, H_3^e, W_3^ϕ, H_3^ϕ). Les matrices de transformation `fixes` T sont retirées des notations lorsqu'elle sont égales à la matrice identité.

Mélange à séparer Le premier signal est le mélange principal à séparer ($m = 1$). Il est composé d'une source de parole V_1, une source de musique V_3 et un bruit additif

9. http ://speech-demos.gforge.inria.fr/source_separation/taslp2015/index.html

V_5 :

$$\begin{aligned} V^1 &= V_1 + V_3 + V_5 \\ &= W_1^e H_1^e \odot W_1^\phi H_1^\phi + W_3^e H_3^e \odot W_3^\phi H_3^\phi + W_5 H_5. \end{aligned} \tag{6.25}$$

Référence de parole Le second mélange est composé de la référence de parole V_2 uniquement :

$$V^2 = V_2 = W_1^e H_2^e \odot T_{12}^{f\phi} W_1^\phi T_{12}^{d\phi} H_1^\phi T_{12}^{t\phi}. \tag{6.26}$$

Pendant les étapes algorithmiques *NMPcF* et/ou *GEM*, H_1^e et H_2^e sont estimées séparément afin de modéliser les différences d'intonation et de *pitch* entre les deux locuteurs. À l'inverse, les matrices du filtre W_1^ϕ et H_1^ϕ sont estimées conjointement afin de modéliser un contenu phonétique similaire, puisque les deux signaux de parole contiennent les mêmes phonèmes. $T_{12}^{t\phi}$ modélise l'alignement temporel entre les deux phrases. $T_{12}^{f\phi}$ est initialisée de sorte à rester diagonale et modélise à la fois l'égalisation fréquentielle et la différence entre les deux locuteurs. $T_{12}^{d\phi}$ modélise également la différence entre locuteurs. Un exemple d'une telle décomposition spectrale est illustrée dans la Figure 6.2. Ce modèle est similaire à celui présenté dans [106], où un filtre invariant en temps modélise les déformations fréquentielles de V_2, cependant ici une matrice de transformation $T_{12}^{d\phi}$ pour le dictionnaire du filtre est en plus utilisée. Les références de parole supplémentaires sont modélisées de la même façon.

Référence de musique Le troisième signal est composé de la référence de musique V_4 similaire à V_3, et d'un bruit V_6 :

$$V^3 = V_4 + V_6 = W_3^e H_3^e T_{34}^{te} \odot W_3^\phi H_3^\phi T_{34}^{t\phi} + W_6 H_6. \tag{6.27}$$

T_{34}^{te} et $T_{34}^{t\phi}$ modélisent l'alignement des spectres entre les deux sources de musique. Les références de musique supplémentaires sont modélisées de la même façon.

Les références de musique V_4 utilisées ici sont des signaux potentiellement très similaires à V_3. D'autres modèles sont proposés dans le chapitre 7 pour le cas de la séparation de signaux communs.

6.4.3 Initialisation des paramètres

Je détaille ici les étapes algorithmiques *Init* et *NMF* définies dans la partie 6.2.3. Le paramètre de pondération $\lambda^{m'}$ est fixé à $\frac{NF}{N'F'}$ pour compenser les différences de durée entre les exemples. Les dictionnaires de composantes spectrales de l'excitation (W_1^e, W_3^e) sont `fixés` comme un ensemble de composantes harmoniques calculées comme décrit dans la partie 6.3.2. Les matrices de synchronisation $T_{12}^{t\phi}$, T_{34}^{te}, et $T_{34}^{t\phi}$ sont initialisées par des chemins d'alignement obtenus par DTW (calculés à l'aide de [54]) sur les séquences de MFCC (calculés à l'aide de [55]) pour les sources de parole et de chroma pour les sources de musique. En se basant sur [106], le chemin d'alignement est autorisé à varier à l'intérieur d'une région élargie proche du chemin estimé (plus de détails sont donnés dans [105]). Les signaux étant déformés et/ou bruités, ce chemin élargi est en

	rapport voix/musique : −6 dB						rapport voix/musique : 12 dB					
	voix			musique			voix			musique		
	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
Init + NMPcF	-0,6	2,3	-2,1	5,9	11,7	8,3	3,6	4,9	29,2	-6,8	6,8	-5,5
Init + NMF + Plain-NMF	1,8	1,0	8,9	9,2	13,1	12,4	5,1	7,1	23,8	-3,9	3,1	-1,5
Init + NMF + NMPcF	**2,1**	2,9	8,1	**9,2**	11,6	17,7	6,0	8,7	24,6	**0,5**	2,7	3,9

Tableau 6.2 – Moyennes des performances de séparation voix/musique (dB) pour différentes combinaisons d'étapes algorithmiques dans le cas de l'utilisation d'une référence de musique et d'aucune référence de voix.

plus pondéré par la matrice de similarité (2.9) correspondant à la DTW afin d'éviter les erreurs grossières d'initialisation. La matrice de transformation spectrale $T_{12}^{f\phi}$ est initialisée par la matrice identité. Les autres matrices (H_1^e, H_2^e, W_5, H_5, W_6, H_6, W_1^ϕ, H_1^ϕ, H_3^e, W_3^ϕ, H_3^ϕ) sont initialisées par des valeurs aléatoires.

L'étape *NMF* peut ensuite être appliquée séparément aux différents mélanges de référence (6.26) et (6.27), où les matrices `partagées` (W_1^ϕ, H_1^ϕ, H_3^e, W_3^ϕ, H_3^ϕ) et les paramètres `libres` (H_2^e, W_6, H_6) sont mis à jour alors que les matrices $T_{12}^{f\phi}$, $T_{12}^{d\phi}$, $T_{12}^{t\phi}$, T_{34}^{te}, et $T_{34}^{t\phi}$ ne le sont pas. W_6 et H_6 sont ensuite réinitialisées par des valeurs aléatoires avant que les étapes *NMPcF* et/ou *GEM* soient appliquées.

6.4.4 Combinaisons algorithmiques

La première expérience a pour but d'évaluer l'effet de l'initialisation par NMF (étape *NMF*) dans le cas d'une seule référence de musique sans référence de parole. Le nombre d'itérations est fixé à 10 pour les étapes *NMF*, *Plain-NMF*, et *NMPcF*. Les performances de séparation sont évaluées en terme de SDR, de SIR et de SAR [174]. Les résultats sont présentés dans le Tableau 6.2. Les meilleurs SDR sont indiqués en gras pour chaque colonne.

On observe une amélioration notable (au moins 2,5 dB) lorsque l'étape *NMF* est appliquée au préalable par rapport à l'étape *NMPcF* seule. L'utilisation de l'étape *NMPcF* au lieu de l'étape *Plain-NMF* mène ensuite à une autre amélioration lorsque la source avec une référence (ici la musique) est en arrière-plan (−12 dB). Un comportement similaire est observé dans les expériences de la partie 6.5.

Le nombre d'itérations pour l'étape *GEM* qui est connue pour nécessiter plus d'itérations est fixé à 100. L'ensemble des résultats incluant l'étape *GEM* est présenté en Annexe D.3.1 et discuté dans la partie 7.3.2.

6.4.5 Multiples références pour une même source

Des expériences complémentaires sur l'effet du nombre de références pour une même source, c'est-à-dire plusieurs références j' pour une seule source j, ont ensuite été conduites. Le nombre de références de parole (resp. de musique) varie de 0 à 3 (resp. de 0 à 2). Les performances de séparation sont également évaluées pour une expérience

Nombre de références de parole	Nombre de références de musique	rapport voix/musique : −6 dB						rapport voix/musique : 12 dB					
		voix			musique			voix			musique		
		SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
1	0	2,1	5,9	3,7	7,7	11,9	13,8	8,7	11,3	19,7	**-2,2**	3,5	-1,5
2	0	2,3	6,1	3,9	7,9	12,4	13,3	8,6	11,1	19,9	-2,5	3,8	-1,9
3	0	**2,8**	5,7	4,7	**8,3**	12,6	13,6	**9,2**	11,7	20,6	**-2,2**	4,1	-1,6
0	1	2,1	2,9	8,1	**9,2**	11,6	17,7	6,0	8,7	24,6	0,5	2,7	3,9
1	1	4,6	6,0	9,9	8,0	9,6	18,9	**13,3**	14,5	26,2	**1,6**	3,4	6,7
2	1	4,9	6,2	10,2	8,5	10,4	19,4	12,2	13,5	25,5	0,9	3,4	6,5
3	1	**5,0**	6,3	10,5	8,6	10,5	19,4	12,1	13,4	25,4	**1,6**	3,4	6,3
0	2	1,5	3,0	4,9	8,4	11,8	14,4	4,8	7,5	25,3	-2,2	3,9	-1,0
1	2	4,1	6,1	8,5	8,1	11,0	16,2	**10,4**	12,2	26,2	-0,6	3,4	1,2
2	2	**4,6**	6,3	9,4	8,3	11,1	16,5	10,3	12,2	26,2	-0,9	3,6	1,3
3	2	**4,6**	6,2	9,5	**8,5**	11,3	16,8	10,0	11,8	25,7	**-0,5**	3,7	1,2
RbWF	*RbWF*	3,0	4,1	7,2	9,0	13,3	12,0	5,5	7,5	24,8	-6,5	6,2	-5,4

Tableau 6.3 – Moyennes des performances de séparation voix/musique (dB) pour différents nombres de références de parole et de musique. Seulement 10 itérations de *NMF* et de *NMPcF* sont appliquées. Les meilleurs SDR sont indiqués en gras.

de base sans aucune modélisation des déformations (sauf l'alignement temporel), que je nomme ci-après *Reference-based Wiener Filter (RbWF)*. Le Tableau 6.3 réunit l'ensemble des résultats.

On peut souligner que l'utilisation de multiples références de parole mène à de meilleurs résultats, en particulier lorsque la source de parole est en arrière-plan (de l'ordre de 0,5 dB). À l'inverse, l'utilisation de deux références de musique conduit à des résultats en moyenne inférieurs ou égaux. Le fait que les références de musique considérées ici contiennent des sources additionnelles nuisibles V_6 dans (6.27) d'une forte variabilité peut être une explication.

D'une façon générale, l'ajout de références supplémentaires semble améliorer la séparation lorsque les nouvelles références apportent de l'information complémentaire. Dans le cas de la musique, deux références distinctes sont utiles lorsque le chevauchement avec les sources nuisibles est différent d'une référence à l'autre. De même, dans le cas de la parole, chaque nouvelle référence (qu'elle soit prononcée par le même locuteur ou non) apporte des occurrences de phonèmes potentiellement plus similaires à celle de la source cible que les références déjà existantes. Toutefois, détenir au moins une référence pour chaque source cible reste le point le plus crucial, même si la référence est très déformée.

L'expérience de base *Reference-based Wiener Filter (RbWF)* est une méthode de séparation non itérative qui se base uniquement sur un filtre de Wiener adaptatif et un alignement temporel des références et du mélange. Une fois alignés au mélange, les spectres de puissance observés des références (V^2 et V^3) sont choisis comme « modèles » pour les sources :

$$V_1 = V^2 T_{12}^t \tag{6.28}$$

$$V_3 = V^3 T_{34}^t, \tag{6.29}$$

avant que l'on applique le filtre de Wiener (3.15). Cette méthode donne une idée de la qualité des références. Les résultats montrent une diminution significative de performance par rapport au cas avec une référence par source, alors que le résultat pour la musique en avant-plan est comparable. Ceci s'explique par la relative bonne qualité des références de musique et le fait que l'alignement soit réussi lorsque la musique prédomine. Toutefois, l'approche proposée démontre un réel avantage lorsque la source ne prédomine pas. Cette expérience de base devrait certainement pouvoir profiter de mesures de similarité plus robustes comme celles développées dans le chapitre 5.

6.5 Séparation de musique guidée par des reprises multi-pistes

Dans le cas de la musique, la séparation de sources audio cherche à fournir les signaux pour chaque instrument ou voix. Cette dernière expérience s'intéresse plus précisément à la tâche de séparation de musique guidée par des reprises multi-pistes [75]. Une reprise de musique est une réplique d'un morceau original avec des différences dues par exemple à l'interprétation artistique, à des changements d'instruments ou de chanteurs ou à une nouvelle structure du morceau. Les versions multi-pistes de ces reprises sont plus facilement accessibles que celles des morceaux originaux et, contrairement à ce qu'on peut imaginer, elles sont généralement proches des pistes originales (pour des raisons commerciales) ce qui les rend intéressantes pour la séparation guidée. En effet, les pistes séparées de la reprise fournissent un moyen simple et précis d'effectuer l'initialisation [75]. Le nombre de pistes est le même que le nombre de sources cibles, chaque piste étant utilisée comme référence pour la source correspondante.

Dans cette partie, plutôt que d'utiliser les différentes pistes de la reprise uniquement pour l'initialisation, elles sont également utilisées pour contraindre le modèle spectral de chaque source. De plus, bien que les pistes de la reprise soient musicalement fidèles à l'original, des déformations à l'échelle du signal existent entre les sources originales et celle de la reprise. Ces déformations sont suffisamment significatives pour ne pas être ignorées. Ici, différentes configurations de déformations (comme formalisé dans la partie 6.1) sont testées. Le modèle optimal de déformation est ensuite sélectionné pour chaque type de source (voix, basse, batterie, guitare...). Enfin, des expériences préliminaires sont menées sur des références et des mélanges multicanaux.

6.5.1 Données et paramètres généraux

De façon à pouvoir comparer les résultats, le même jeu de données et de paramètres que dans [75] est utilisé. Afin de permettre l'évaluation de la séparation, les pistes séparées des morceaux originaux et des reprises sont disponibles. Ces pistes sont également utilisées dans la configuration inverse, c'est-à-dire en considérant l'original comme référence et vice-versa. Les mélanges mono et stéréo sont produits par un ingénieur du son [75]. Je dresse ci-après une liste des paramètres qui diffèrent de [75].

Les exemples de 30 secondes sont choisis de façon différente que dans [75] et sont ty-

Titre	**Noms des pistes**
I Will Survive	Basse, Batterie, Cordes, Cuivres, Guitare, Voix.
Pride et Joy	Basse, Batterie, Guitare, Voix.
Rocket Man	Basse, Batterie, Chœurs, Piano, Voix, Autres.
Walk this Way	Basse, Batterie, Guitare, Voix.

Tableau 6.4 – Base de données de reprises multi-pistes.

piquement composés de la moitié d'un couplet et de la moitié d'un refrain. Les pistes des quatre morceaux utilisés sont listées dans le Tableau 6.4. Des exemples sont disponibles en ligne [10]. On utilise 50 itérations pour les étapes *NMF* et *NMPcF* au lieu de 500 [75] et 10 itérations pour l'étape *GEM* au lieu de 500 [75]. Le nombre de composantes de NMF est conservé à 50. Par souci de clarté, le cas mono-canal est tout d'abord étudié pour montrer les effets des différents modèles de déformation, puis les mélanges stéréo sont traités par l'algorithme *GEM* présenté dans la partie 6.2.2.

6.5.2 Modèles testés

Dans ce scénario, le mélange à séparer est un morceau de musique orignal et les signaux de référence sont les différentes pistes de la reprise de ce morceau. Chaque signal de référence est associé à une source cible dans le mélange à séparer. Ici, nous considérons que $\mathbf{x}^1(t)$ est l'original à séparer et que $\mathbf{x}^m(t)$ pour $m > 1$ sont les différentes pistes de la reprise supposées ne contenir qu'une seule source.

Chaque V_j est décomposé comme une NMF simple $V_j = W_j H_j$. On considère uniquement les matrices de transformation fréquentielle et de dictionnaire qui sont maintenant notées $T^f_{jj'} \in \mathbb{R}_+^{F\times F}$, et $T^d_{jj'} \in \mathbb{R}_+^{D\times D}$. Comme les pistes des deux versions sont suffisamment alignées temporellement, aucune matrice T^t n'est utilisée car elles induisent des effets de lissage indésirables. Ainsi, la source correspondante est modélisée en utilisant l'équation (6.7) :

$$V_{j'} = T^f_{jj'} W_j T^d_{jj'} H_j. \tag{6.30}$$

On peut remarquer que cette formulation laisse la possibilité de placer ces matrices de transformation soit dans le modèle de référence ($j \in \mathcal{J}^1$ et $j' \in \mathcal{J}^{m'}$, $m' \neq 1$) soit dans le modèle de la source ($j' \in \mathcal{J}^1$ et $j \in \mathcal{J}^{m'}$, $m' \neq 1$). L'inversion des j et j' permet par exemple de ne pas perturber l'initialisation effectuée par l'étape *NMF*, les matrices T^f et T^d étant mal initialisées. Se référer aux Tableaux 6.6 et 6.7 de la partie 6.5.3 pour des cas concrets. Pour T^f ou T^d, deux initialisations sont possibles :

- `Diag` : une matrice identité,
- `Full` : la somme d'une matrice identité et d'une matrice dont les éléments sont tirés aléatoirement en suivant une loi normale rectifiée [11].

10. http ://speech-demos.gforge.inria.fr/source_separation/icassp2015/

11. Plusieurs initialisations de T^d ont été testées en partant de la matrice identité et en changeant le poids des termes non diagonaux. L'initialisation `Full` résulte de ces tests.

	Modèle conjoint		**SDRI**
	Source	Référence	**moyen**
Init + *NMF* + *Plain-NMF* (Résultats dans [75])	WH		8,98
Init + *NMF* + *Plain-NMF* (Reproduction de [75])	WH		8,74
Init + *NMF*			10,06
Init + *NMF* + *NMPcF*	WH	WH	10,27

Tableau 6.5 – SDRI (dB) moyens par rapport à une précédente étude [75].

Les paramètres W et H sont initialisés aléatoirement avant d'être mis à jour pour s'ajuster au signal de référence durant l'étape *NMF* de la même façon que dans [75]. Lorsque l'on traite les données stéréo, l'étape *GEM* est utilisée. Le paramètre spatial de rang plein est alors initialisé au préalable en utilisant la référence pour chaque source comme dans [75].

Du fait que les paramètres spectraux soient mis à jour par MU, je rappelle que les paramètres initialisés à zéro le resteront au cours des itérations. Ainsi, une matrice initialisée à l'identité restera diagonale. De plus, les matrices T ne sont pas présentes pendant l'étape préliminaire *NMF*.

6.5.3 Résultats

La qualité des sources estimées est évaluée en terme d'amélioration du SDR (SDR *improvement* ou SDRI) qui est la différence entre le SDR [174] de sortie et le SDR d'entrée. Le SDR d'entrée est défini comme le rapport de puissance entre la source à estimer et le reste des sources présentes dans le mélange à séparer. Il est donné pour chaque source dans les Tableaux 6.7 et 6.8. Les exemples sélectionnés mènent à un SDR d'entrée de $-8{,}44$ dB au lieu de $-7{,}60$ dB dans [75].

6.5.3.1 Comparaison avec une précédente étude

Dans [75], les signaux multi-pistes de la reprise sont uniquement utilisés pour initialiser les paramètres des sources W et H (*Plain-NMF*). Ici, les paramètres sont également partagés entre les sources et les références, ainsi les signaux de référence sont aussi utilisés durant l'estimation globale (*NMPcF*) des paramètres. Les résultats sont donnés par le Tableau 6.5.

Tout d'abord, l'expérience dans [75] a été reproduite avec les différences présentées dans la partie 6.5.1. Un SDRI équivalent à [75] est obtenu dans le cas où les paramètres ne sont pas partagés (8,74 dB au lieu de 8,98 dB). Cette configuration mène en fait à une importante diminution du SDRI moyen par rapport à ce qui est obtenu si les sources sont directement reconstruites après l'étape préliminaire *NMF* (10,06 dB). Une explication peut être le niveau important de similarité entre les pistes de la reprise et de l'original, comme montré par les résultats de la méthode *Reference-based Wiener Filter*

Init		Modèle conjoint		**SDRI**
T^f	T^d	Source	Référence	**moyen**
		WH	WH	10,27
`Full`		WH	T^fWH	10,08
`Full`		T^fWH	WH	9,23
`Diag`		T^fWH	WH	10,09
`Diag`		WH	T^fWH	**10,35**
	`Diag`	WT^dH	WH	9,25
	`Diag`	WH	WT^dH	9,88
	`Full`	WH	WT^dH	9,79
	`Full`	WT^dH	WH	**10,64**

Tableau 6.6 – SDRI (dB) moyen pour différentes configurations.

(RbWF) présentés dans le Tableau 6.7 [12]. Inversement, partager les paramètres durant l'estimation finale *NMPcF* garantit de ne pas s'éloigner de ce point pertinent de départ tout en se rapprochant d'une solution qui s'ajuste aux pistes de l'original. Dans notre cas, une amélioration marginale est observée (10,27 dB).

Ces premiers résultats montrent la forte similarité entre chaque piste originale et la piste correspondante dans la reprise. Dans ce cas, partager W et H durant l'estimation conjointe (*NMPcF*) est la méthode la plus pertinente même en ne considérant aucune déformation.

6.5.3.2 Modèle de déformation

Nous allons ensuite analyser si les matrices de transformation sont plus utiles dans le modèle de la référence ou de la source. La comparaison des différentes initialisations des matrices de déformation fréquentielle ou de dictionnaire est également effectuée. Les résultats complets sont exposés dans le Tableau 6.6.

Les valeurs en gras indiquent une amélioration par rapport à un modèle conjoint totalement partagé (10,27 dB). On peut remarquer que dans ces deux cas, le nombre de coefficients non nuls Z dans T est du même ordre de grandeur ($Z = D^2 = 2500$ pour les matrices `Full` T^d et $Z = F = 1025$ pour les matrices `Diag` T^f), alors que dans les autres cas Z varie de $D = 50$ à $F^2 \approx 10^6$.

On observe également que dans presque tous les cas les SDRI sont toujours plus élevés lorsque les matrices T sont placées dans le modèle de référence. Ceci peut être expliqué par le fait que la reconstruction du signal final est basée sur le modèle de la source, et que les matrices T peuvent induire des changements abrupts. À l'inverse, ajouter une matrice T `Full` dans le modèle de la référence fausserait la sortie de l'étape préliminaire *NMF*. En effet, le produit de W_j et H_j estimé pendant cette étape cherche

12. Ces résultats ont été obtenus en suivant la définition de *RbWF* donnée dans la partie 6.4.5 à l'exception qu'ici aucun alignement temporel n'a été opéré.

à s'ajuster au signal de référence. En conclusion, il est difficile de distinguer quel effet est prédominant, d'autant plus que le nombre de coefficients non nuls a également un impact.

On peut retenir qu'il est important de ne pas déformer la sortie de l'étape *NMF* par l'ajout de matrices T. Par exemple dans le Tableau 6.6, c'est le cas pour les valeurs en gras.

Coûts algorithmiques Se pose également la question du coût algorithmique des modèles présentés. L'étape d'initialisation qui est utilisée avant chaque estimation globale nécessite 0.4 millions d'opérations élémentaires [13] par itérations. Il faut rajouter 0.4 millions d'opérations par itérations de *Plain-NMF* et 0.8 millions d'opérations par itérations de *NMF* conjointe. Les itérations incluant des matrices T^d `Full` ont quant à elles besoin de 1.2 millions d'opérations alors que celles incluant des matrices T^f `Diag` 4.6 millions [14].

On peut noter que pour le même ordre de grandeur de paramètres non nuls dans les matrices de déformation (2500 pour T^d et 1025 pour T^f) le nombre d'opérations est nettement inférieur pour T^d. Les propriétés de réduction de dimension de la NMF se retrouvent ainsi également dans l'estimation des matrices de déformations de l'axe des composantes du dictionnaire. Ces valeurs permettent de mettre en relief les gains en qualité de séparation apportés par ces modèles en vue d'une application réelle où l'ingénieur du son pourrait favoriser le temps de traitement à la qualité de traitement.

6.5.3.3 Modèle spécifique de source

Dans cette dernière expérience sur les signaux mono-canal, les SDRI pour chaque type de source sont donnés dans le Tableau 6.7 pour différentes configurations. La combinaison de T^d et T^f donne des résultats intéressants, en particulier pour la batterie et les basses. De plus, pour chaque source, on observe des différences claires entre modèles alors que le SDRI moyen est le même. Dans l'expérience dénommée « **Best** », les configurations les plus appropriées sont choisies pour chaque type de source (et indiquées en gras dans le tableau), la configuration *Init + NMF* n'étant pas prise en compte car cela empêcherait les sources concernées de bénéficier des mises à jour des autres paramètres durant la *NMPcF*. Les valeurs qui sont aussi en italique ont, quant à elles, été choisies en raison des expériences avec une matrice T^f (`Diag`) dans le modèle de la source qui ont montré des résultats prometteurs pour ces deux sources. Le résultat (10,92 dB) montre l'intérêt de définir des modèles spécifiques en fonction du type de source. On a aussi pu observer qu'utiliser des matrices de transformation uniquement pour une seule source conduit à une importante diminution de la qualité de séparation pour cette source. Ceci est certainement dû à l'algorithme d'estimation utilisé. Définir des modèles spécifiques aux sources est donc intéressant mais le nombre de paramètres doit être bien réparti entre les différents modèles.

13. Addition, soustraction, multiplication et division.
14. 9.7 millions pour les itérations avec des matrices T^f `Full`.

Du fait que les modèles ont été sélectionnés en fonction du résultat optimal sur l'ensemble de test, cette dernière expérience n'est pas représentative d'un scénario non supervisé. Ce scénario est cependant réaliste dans notre contexte où un utilisateur expert devrait être capable de sélectionner les meilleures modèles en écoutant les résultats. Des expériences supplémentaires sont nécessaires pour comprendre quels sont les paramètres qui doivent lui être accessibles, éventuellement en fonction de son niveau d'expertise.

6.5.3.4 Résultats multicanaux

La configuration stéréo exposée dans [75] apporte une amélioration importante en terme de SDRI par rapport à la configuration mono (de 8,98 dB à 10,05 dB). L'explication alors avancée suggérait que l'amélioration globale était due en particulier à quelques instruments ayant une position spatiale spécifique (comme la guitare).

Ici, le modèle sans déformation montre une légère amélioration (de 10,27 dB à 10,41 dB) lorsque 10 itérations supplémentaires de l'étape *GEM* sont appliquées et des résultats similaires sont observés pour les modèles avec déformations. Cependant, on n'observe pas autant d'amélioration que dans [75] et pas non plus de grosses différences entre les sources spatialisées ou centrées. L'algorithme EM requiert généralement plus d'itérations qui n'étaient pas envisageables en raison du temps de calcul. En effet, ici chaque référence est considérée comme un mélange ce qui nécessite une étape E spécifique pour chaque référence dans l'algorithme GEM présenté dans la partie 6.2.2.

6.6 Conclusion

Dans ce chapitre, nous avons présenté un cadre général pour l'utilisation de signaux de référence pour la séparation de sources. Ce modèle est assez général pour prendre en compte différents types de références audio et pour s'adapter à leurs potentielles déformations dans le domaine fréquentiel et/ou temporel. Après avoir présenté un scénario élémentaire avec des exemples *pitch-shiftés* artificiellement, deux scénarios réalistes ont été explorés au cours d'expériences approfondies : la séparation de mélanges voix/musique dans le contexte de bandes-son de films et la séparation de musique guidée par les différentes pistes d'une reprise.

L'utilisation d'une ou plusieurs références pour une source donnée améliore généralement la qualité sonore de la source estimée (de 9 à 15 dB SDRI). De plus, les expériences montrent qu'avoir au moins une référence par source est primordial.

Perspectives Les résultats obtenus pour les différentes configurations du modèle général seront certainement utiles pour d'autres scénarios de séparation de sources audio guidée par référence.

Une application possible est l'utilisation de références sélectionnées par un utilisateur de même que des modèles construits par un utilisateur. Une perspective plus générale serait l'étude de procédés automatisant le choix de la meilleure configuration du modèle pour une référence et une source cible données.

Modèle conjoint		**Instruments**	**Moyenne**	**Basse**	**Batterie**	**Guitare**	**Voix**	**Chœurs**	**Autres**	**Piano**	**Cuivres**	**Cordes**
		Nombre de pistes	20	4	4	3	4	1	1	1	1	1
Source	Référence	SDR en entrée	-8,44	-7,42	-7,17	-9,98	-4,18	-12,34	-9,75	-12,48	-18,64	-10,55
		Reference-based Wiener Filter	9,93	9,27	8,90	9,59	9,42	13,40	9,88	10,34	16,30	9,51
		Init + NMF	10,06	9,33	9,02	9,71	9,60	13,70	9,79	10,80	16,25	9,80
WH		*Init + NMF + Plain-NMF* [75]	8,74	6,94	8,95	8,53	8,14	11,36	9,66	10,51	11,54	10,11
WH	WH	*Init + NMF + NMPcF*	10,27	9,26	9,28	9,82	10,24	13,23	10,27	11,11	15,67	10,62
WH	T^fWH	*Init + NMF + NMPcF*	10,35	9,39	9,29	9,69	10,22	**13,85**	***9,98***	***12,08***	15,68	**10,71**
WT^dH	WH	*Init + NMF + NMPcF*	10,64	9,23	9,94	**10,80**	**10,73**	13,01	10,07	11,36	15,80	10,61
WT^dH	T^fWH	*Init + NMF + NMPcF*	10,66	**9,88**	**10,48**	9,41	10,44	12,74	10,01	12,24	**16,66**	10,11
Best		*Init + NMF + NMPcF*	10,92	9,66	10,58	10,35	10,85	13,73	10,74	11,91	14,91	11,71
WH	WH	*Init* + Oracle	15,48	14,48	15,35	15,36	15,29	16,21	13,27	16,22	21,58	15,74

Tableau 6.7 – SDRI (dB) moyen pour la séparation d'enregistrements de musique en utilisant les différentes pistes d'une reprise comme références. Les valeurs en gras indiquent les modèles choisis pour l'expérience **Best**.

	Modèle conjoint		**Instruments**	**Moyenne**	**Basse**	**Batterie**	**Guitare**	**Voix**	**Chœurs**	**Autres**	**Piano**	**Cuivres**	**Cordes**
			Nombres de pistes	20	4	4	3	4	1	1	1	1	1
	Source	Référence	SDR en entrée	-8,44	-7,42	-7,17	-9,98	-4,18	-12,34	-9,75	-12,48	-18,64	-10,55
Mono	WH	WH	*Init + NMF + NMPcF*	10,27	9,26	9,28	9,82	10,24	13,23	10,27	11,11	15,67	10,62
Mono	T^fWH	WH	*Init + NMF + NMPcF*	10,09	8,79	9,22	9,88	9,78	12,29	10,47	12,94	14,66	10,65
Mono	WT^dH	WH	*Init + NMF + NMPcF*	10,64	9,23	9,94	10,80	10,73	13,01	10,07	11,36	15,80	10,61
Stereo	WH	WH	*Init + NMF + NMPcF + GEM*	10,41	8,99	9,12	10,05	9,98	15,31	10,96	11,95	16,79	10,76
Stereo	T^fWH	WH	*Init + NMF + NMPcF + GEM*	10,16	8,64	8,86	8,87	10,02	15,50	10,90	12,78	17,27	10,05
Stereo	WT^dH	WH	*Init + NMF + NMPcF + GEM*	10,55	9,00	9,73	10,05	10,50	13,47	10,24	11,70	17,83	10,62

Tableau 6.8 – SDRI (dB) moyen pour la séparation d'enregistrements mono et stéréo de musique en utilisant les différentes pistes d'une reprise comme références.

Les effets de la surparamétrisation, c'est-à-dire trop de degrés de liberté pour le modèle conjoint ou les matrices de transformations, mériteraient d'être étudiés dans des conditions plus contrôlées ou encore de façon théorique. De même, les nombres minimum d'échantillons ou d'observations nécessaires à l'estimation de ces modèles ne sont pour le moment pas déterminés.

Chapitre 7

Modèle d'alignement fin pour la séparation de signaux communs

Le chapitre précédent a présenté un modèle permettant de rendre compte d'une grande variété de déformations entre une source cible et sa source de référence. Sans parler des autres sources ou du bruit, ces déformations pouvaient affecter les composantes de la NMF, l'axe des fréquences ou l'axe temporel. Les signaux correspondants à la source cible et à la source de référence pouvaient alors être très différents du point de vue de la forme d'onde. Cependant, il existe des situations où ces signaux sont identiques ou très similaires au niveau du signal (« signaux communs »).

L'application visée dans ce chapitre reste la séparation de bandes-son de films, ici avec l'utilisation de références de musique. Dans le cadre de l'approche SPORES, les références peuvent être identifiées par un algorithme de recherche de motifs et dans le cas particulier de signaux communs elles peuvent être :

— des motifs musicaux ou effets sonores dont le signal a été réutilisé ailleurs (par exemple dans le même film) mélangé à d'autres sources,
— des bandes originales de la musique,
— ou des bandes-son complètes dans d'autres langues.

Le modèle proposé dans le chapitre précédent n'est alors pas adapté car il autorise des déformations trop importantes. De plus, ce type de signaux de référence est susceptible d'apporter des informations beaucoup plus précises sur les sources cibles que des signaux de référence qui ne partagent comme propriétés que le moyen de production du signal (instrument, locuteur) ou le contenu symbolique du signal (phonème, note). Il apparait donc opportun de s'intéresser à une modélisation plus fine des déformations entre ces signaux.

Ce chapitre s'intéresse en particulier au cas d'enregistrements ayant subi des déformations analogiques, dues notamment à des différences de tension de bande. Les déformations temporelles induites sont variables au cours du temps et nécessitent un recalage de phase. Sprechmann *et al.* [163] ont proposé l'utilisation d'une DTW (voir partie 2.2) à l'échelle des échantillons pour le bruit impulsif à partir de plusieurs copies de grammophones. Cette technique est cependant plus proche de celles utilisées en

inpainting [2] ou en *declipping*, et n'est pas adaptée à la présence de bruits plus diffus dans le temps. Nous chercherons typiquement à estimer des délais avec une précision fractionnaire, c'est-à-dire inférieure au pas d'échantillonnage ($< 10^{-4}$ secondes). Il est donc nécessaire d'envisager des techniques d'alignement beaucoup plus fines que celles de type DTW (que ce soit sur les échantillons ou sur des trames de STFT), mais toujours capables d'estimer un alignement temporel variable au cours du temps (comme la DTW).

Pour cela nous nous inspirons des techniques existantes de séparation de signaux communs et d'estimation de délais d'arrivée. Après un bref état de l'art de ces techniques, ce chapitre propose un nouveau modèle et l'algorithme correspondant (*GEM-PHAT*) permettant la prise en compte de ces deux caractéristiques. Cet algorithme est évalué pour la séparation de mélanges voix/musique guidée par des références de musique déphasées artificiellement ou par des références de musique provenant du même film. L'algorithme sera comparé sur ces deux tâches à l'algorithme *GEM* déjà présenté dans la partie 6.2 qui est similaire mais qui n'inclut pas de recalage de phase comme *GEM-PHAT*. Enfin l'algorithme sera testé pour la séparation de la musique dans les bandes-son de films, guidée par une version dans une autre langue.

7.1 État de l'art

Le problème traité dans ce chapitre est assez atypique et n'a encore jamais été exploré avec les contraintes précédemment énoncées. L'état de l'art que je vais présenter maintenant regroupe différentes approches de la littérature pouvant paraitre assez éloignées les unes des autres par la nature des techniques utilisées (NMPcF, modélisation multicanale, *PHAse Transform* (PHAT), DTW) ou les applications visées (séparation, localisation, restauration). Elles forment, malgré leur hétérogénéité, le contexte scientifique de notre problème et de l'approche que j'ai envisagée pour sa résolution.

7.1.1 Séparation de signaux communs

La partie 3.4 a donné la définition suivante des problèmes de séparation de signaux communs : « la même source apparait dans plusieurs mélanges », ainsi qu'une catégorisation parmi les approches de séparation guidées par signal de référence. Cette catégorisation n'est pas unique et les canaux des mélanges peuvent également être considérés [117] comme contenant des signaux communs lorsqu'il s'agit d'enregistrements réels provenant par exemple d'antennes de microphones. De plus, l'extraction d'un fond musical répétitif peut aussi s'apparenter à de la séparation de signaux communs si l'on suppose que la redondance n'est pas produite par répétition musicale mais par répétition des signaux eux-mêmes. Ce n'est cependant pas l'hypothèse courante de REPET [144] ou de [117].

La principale application de la séparation de sources communes est l'extraction de musique dans des films à partir de plusieurs versions en différentes langues [29,109,113]. La même musique est alors considérée comme présente dans les différents signaux à l'inverse des sources de parole. Les autres applications possibles sont la restauration

audio [163], le rehaussement collaboratif [101] et plus généralement la séparation multicanale.

Je détaille ci-après les techniques existantes que l'on peut regrouper en deux catégories principales qui sont d'une part les approches spectrales et d'autre part les approches spatiales.

Approches spectrales La plupart des approches spectrales peuvent être représentées par le cadre proposé dans le chapitre 6. Dans l'approche dénommée « *Convolutive Common Nonnegative Matrix Factorization* » par Leveau *et al.* [109], des matrices W et H sont partagées par les sources communes et la déformation est modélisée par différents gains pour chaque canal. De même, l'approche dénommée « *Probabilistic Latent Components Sharing* » [101] est la version cofactorisée de la PLCA. Les probabilités $P(f|k)$ et $P(n|k)$ de l'équation (3.17) sont alors identiques pour les différentes contributions de la source commune dans les différents mélanges (ou canaux).

Ce premier groupe d'approches se focalise sur la modélisation du spectre de puissance et ne prend pas en compte la phase des signaux ce qui représente une limitation pour la résolution du problème traité dans ce chapitre.

Approches spatiales Les approches spatiales telles que [29, 113] peuvent s'avérer suffisantes dans des cas simples de positionnement de la musique par rapport aux sources de voix situées dans un seul canal. Mais elle ne prennent pas en compte la phase.

L'utilisation d'un modèle convolutif de rang 1 [131] permet de prendre en compte la phase des signaux contrairement à un simple facteur de gain [109, 113]. Cependant ce modèle reste limité [44] par rapport aux modèles spatiaux de rang plein (équation (3.10)) capables de traiter des cas plus complexes comme la réverbération. Dans ce chapitre (voir partie 7.3.1), je privilégie l'utilisation de ces derniers que ce soit dans l'algorithme *GEM-PHAT* proposé dans la partie 7.2 ou le *GEM* proposé dans le chapitre 6 qui va servir de comparaison.

Les techniques de séparation de sources communes présentées ci-dessus font toutes l'hypothèse que les signaux sont alignés, ou alors qu'il existe un délai global d'alignement qui doit être estimé [113]. Les techniques qui ne prennent pas en compte la phase font l'hypothèse que ce délai ne varie pas au cours du temps de plus d'une fenêtre de STFT. Quant aux techniques de séparation spatiale qui prennent en compte la phase, elles font en plus l'hypothèse que l'alignement en phase des signaux est correct et qu'il ne varie pas au cours du temps. C'est notamment à cette limitation que ce chapitre s'attaque.

7.1.2 Estimation de délais

Nous nous intéressons ci-après à une technique d'estimation de délais nommée GCC-PHAT qui sera par la suite réutilisée par le recalage de phase et par le modèle de source avec délai proposé dans la partie 7.2.

GCC-PHAT L'estimation d'un délai entre deux canaux i_1 et i_2 (*Time Delay Of Arrival* ou TDOA) se fait par maximisation d'une fonction de corrélation croisée géné-

ralisée (*Generalized Cross Correlation* ou GCC). Son calcul est effectué dans le domaine temps-fréquence par transformée de Fourier inverse

$$R_{i_1,i_2,n}(\tau) = \sum_f \psi_f \mathbf{X}_{i_1,fn} \mathbf{X}^*_{i_2,fn} e^{j2\pi f\tau} \tag{7.1}$$

où τ est le délai en secondes [1] et ψ_f un facteur de pondération. La pondération la plus connue est appelée PHAT [103]

$$\psi_f^{\text{PHAT}} = \frac{1}{|\mathbf{X}_{i_1,f}\mathbf{X}^*_{i_2,f}|}. \tag{7.2}$$

Le délai optimal est celui qui maximise la fonction de corrélation croisée.

La pondération PHAT retire l'amplitude du spectre pour ne conserver que sa phase ce qui favorise par moyenne le délai principal. Cette technique est cependant limitée lorsque le niveau de bruit augmente, notamment lorsque la source d'intérêt (celle dont on veut estimer τ) ne domine plus dans assez de bandes de fréquences. Il existe d'autres façons de pondérer la corrélation notamment au sens du maximum de vraisemblance lorsque la puissance du bruit est connue pour chaque bande de fréquence [24].

Cette information permet d'aligner les signaux entre eux sur une période donnée. Elle peut aussi être utilisée pour localiser spatialement une ou plusieurs sources en exploitant les informations des différentes paires de microphones. La technique de localisation SRP-PHAT (*Steered Response Power Phase Transform*) [42] consiste par exemple à sommer les fonctions GCC-PHAT de toutes les paires.

Recalage de phase (puis séparation) Pour les besoins d'un projet annexe à cette thèse, une méthode de base (décrite ci-après) a été développée [2] afin de recaler en phase deux signaux avant leur utilisation par un algorithme de séparation de sources multicanale prenant en compte la phase.

[Confidentiel]

La principale limitation de cette méthode de base est que le recalage des signaux peut être incorrect, notamment lorsque les sources communes ne dominent pas suffisamment dans leurs canaux respectifs. Or, une fois établi, cet alignement ne variera plus au cours des étapes algorithmiques suivantes et les erreurs commises ne seront pas corrigées. Afin de résoudre ce problème, la partie suivante propose un algorithme qui estime conjointement cet alignement et les sources.

7.2 Algorithme GEM-PHAT

Principe : recalage de phase et séparation conjointe J'ai proposé dans la partie 6.2.3 plusieurs étapes algorithmiques (*NMF*, *NMPcF* et *GEM*) dont la combinaison permet d'estimer des modèles de séparation guidée par référence. En particulier, *GEM*

1. Par abus de notations, f désigne à la fois les indices de la STFT et la fréquence en Hertz
2. par Emmanuel Vincent

est une technique de séparation spatiale prenant en compte la phase, ce qui nécessite qu'elle soit alignée.

Dans cette partie, je propose une nouvelle étape algorithmique *GEM-PHAT* capable d'estimer conjointement la séparation et le recalage en phase.

Modèle La formulation du problème de mélange est la même que celle de l'équation (3.26) [3]

$$\mathbf{x}_{fn} = \mathbf{A}_{fn}\mathbf{s}_{fn} + \mathbf{b}_{fn}. \tag{7.3}$$

On décompose la matrice de mélange comme

$$\mathbf{A}_{fn} = [\mathbf{D}_{j,fn}\mathbf{\Lambda}_{j,f}]_{j\in\mathcal{J}}, \tag{7.4}$$

où $\mathbf{\Lambda}_{j,f} \in \mathbb{C}^{I\times R_j}$ modélise les caractéristiques spatiales invariantes dans le temps et $\mathbf{D}_{j,fn} = \text{diag}(1,e^{-2i\pi f\tau_{2,jn}},...,e^{-2i\pi f\tau_{I,jn}}) \in \mathbb{C}^{I\times I}$ modélise les délais entre les canaux au cours du temps. $\tau_{i,jn}$ est le délai en seconde de la source j entre le canal 1 et i sur la trame n. Le modèle de mélange (3.26) devient alors

$$\mathbf{x}_{fn} = \sum_{j\in\mathcal{J}} \mathbf{D}_{j,fn}\mathbf{\Lambda}_{j,f}\mathbf{s}_{j,fn} + \mathbf{b}_{fn}. \tag{7.5}$$

L'algorithme *GEM-PHAT* alterne ensuite les étapes « E-step » et « M-step » suivantes afin de faire croitre la vraisemblance.

E-step Les matrices $\widehat{\mathbf{R}}_{\mathbf{xs}_{fn}}$ et $\widehat{\mathbf{R}}_{\mathbf{s}_{fn}}$ sont respectivement obtenues par les équations (3.28) et (3.29).

M-step L'étape M cherche à faire croitre l'espérance de la log-vraisemblance qui s'exprime comme dans (3.27) :

$$\begin{aligned} Q(\theta,\theta^c) \overset{c}{=} -\sum_{fn}\frac{1}{\sigma_f^2}\text{tr}\left[\widehat{\mathbf{R}}_{\mathbf{x}_{fn}} - \mathbf{A}_{fn}\widehat{\mathbf{R}}_{\mathbf{xs}_{fn}}^H - \widehat{\mathbf{R}}_{\mathbf{xs}_{fn}}{\mathbf{A}_{fn}}^H + \mathbf{A}_{fn}\widehat{\mathbf{R}}_{\mathbf{s}_{fn}}{\mathbf{A}_{fn}}^H\right] \\ - \sum_{j\in\mathcal{J},fn} R_j d_{IS}(\widehat{\xi}_{j,fn}|v_{j,fn}). \end{aligned} \tag{7.6}$$

Dans ce but, les mises à jour des paramètres spectraux restent identiques à (3.35) et cherchent à faire diminuer $\sum_{j\in\mathcal{J},fn} R_j d_{IS}(\widehat{\xi}_{j,fn}|v_{j,fn})$.

Concernant les paramètres spatiaux, la mise à jour (3.33) est modifiée en raison de la nouvelle formulation de $\mathbf{A}_{fn}$ (7.4), mais cherche toujours à faire diminuer Q. De la même façon que dans [136], nous introduisons les matrices

$$\widehat{\mathbf{R}}_{\mathbf{xs}_{j,fn}} = [\widehat{\mathbf{R}}_{\mathbf{xs}_{fn}}(i,r)]_{i=1,r\in\mathcal{R}_j}^{I} \in \mathbb{C}^{I\times R_j} \tag{7.7}$$

$$\widehat{\mathbf{R}}_{\mathbf{s}_{jj',fn}} = [\widehat{\mathbf{R}}_{\mathbf{s}_{fn}}(r,r')]_{r\in\mathcal{R}_j,r'\in\mathcal{R}_{j'}} \in \mathbb{C}^{R_j\times R_{j'}} \tag{7.8}$$

3. ou de l'équation (6.13) mais sans les m mélanges.

définies pour deux ensembles d'indices de sous-sources $\mathcal{R}_j$ et $\mathcal{R}_{j'}$ de taille $R_j = \#(\mathcal{R}_j)$ et $R_{j'} = \#(\mathcal{R}_{j'})$. La matrice $\mathbf{\Lambda}_{j,f}$ est mise à jour comme suit pour chaque source :

$$\mathbf{\Lambda}_{j,f} = \left[\sum_n \mathbf{D}_{j,fn}^H \left\{\widehat{\mathbf{R}}_{\mathbf{x}\mathbf{s}_{j,fn}} - \sum_{\tilde{\jmath}\in\mathcal{J}\backslash j} \mathbf{D}_{\tilde{\jmath},fn}\mathbf{\Lambda}_{\tilde{\jmath},f}\widehat{\mathbf{R}}_{\mathbf{s}_{\tilde{\jmath}j,fn}}\right\}\right]\left[\sum_n \widehat{\mathbf{R}}_{\mathbf{s}_{jj,fn}}\right]^{-1} \tag{7.9}$$

où $\mathcal{J}\backslash j$ dénote l'ensemble des sources dont j est exclu. Le détail du calcul permettant d'obtenir cette mise à jour est donné en Annexe (C.2).

Enfin, la mise à jour des paramètres de délai est effectuée à chaque étape M. Le terme Q de l'équation (7.6) est calculé pour différentes valeurs de délais $\tau_{i,jn}$ faisant partie d'une grille prédéfinie. Les délais choisis sont ceux qui font le plus augmenter l'espérance de la log-vraisemblance à chaque trame. L'ensemble des valeurs de délai testées s'étend sur un quart de fenêtre STFT avec un pas de $\frac{1}{32}$ d'échantillon.

7.3 Expériences de séparation voix/musique

J'ai proposé dans la partie 6.4 un premier modèle pour les références de musique (voir équations (6.25) et (6.27)). Je propose dans cette partie deux autres façons de modéliser ces références de musique en prenant en compte les déformations fines évoquées au début de ce chapitre. Les signaux des références de musique sont considérés comme un second canal du mélange à séparer par les étapes algorithmiques *GEM* ou *GEM-PHAT*. Dans les deux cas, les signaux seront recalés en phase au préalable par l'algorithme de recalage basé sur GCC-PHAT.

Après avoir détaillé l'initialisation de ces deux étapes, différentes combinaisons d'algorithmes seront testées sur la tâche de séparation voix/musique du chapitre 6. L'étape *GEM-PHAT* sera ensuite évaluée sur des exemples synthétiques qui reproduisent des décalages de phase. Enfin, l'utilisation de ces algorithmes sur des bandes-son réelles sera discutée.

7.3.1 Initialisation des paramètres

La prise en compte des références de musique comme un second canal a quelques conséquences sur la modélisation et l'initialisation des paramètres de l'étape *GEM-PHAT* (et *GEM*).

Paramètres spatiaux et de délai En repartant des notations de la partie 6.4, V^3 (6.27) n'est plus utilisé dans ce modèle, et V_3 devient une source de musique unique avec un paramètre spatial `libre` $\Lambda_3 \in \mathbb{C}^{2\times 2}$ qui encode l'amplitude et les différences de phase entre les canaux cible et de référence. Ce paramètre est initialisé par $\Lambda_3 \approx \begin{pmatrix} 1 & -0{,}25 \\ 1 & 0{,}25 \end{pmatrix}$, où la première colonne représente le fait que les deux canaux sont supposés être alignés en temps et en amplitude et la seconde colonne prend en compte les différences résiduelles. Les paramètres de délai $\tau_{2,3n}$ sont eux initialisés à

zéro. Comme les autres sources n'apparaissent que dans un seul canal, leurs paramètres spatiaux resteront `fixés` à $\Lambda_1 = \Lambda_5 = \begin{pmatrix}1\\0\end{pmatrix}$ et $\Lambda_6 = \begin{pmatrix}0\\1\end{pmatrix}$. De même, les matrices de délai de ces sources sont désactivées, c'est-à-dire fixées à $\mathbf{D}_{1,fn} = \mathbf{D}_{5,fn} = \mathbf{D}_{6,fn} = \mathbf{I}$.

Les modèles de l'étape algorithmique *GEM* sont initialisés de la même façon hormis la matrice de délai de la source de musique qui reste inactive et fixe durant l'estimation : $\mathbf{D}_{3,fn} = \mathbf{I}$. Ainsi, on revient dans le cadre de l'algorithme GEM décrit dans la partie 6.2.

Paramètres spectraux H_3^e, W_3^ϕ, et H_3^ϕ sont des paramètres `libres` et les matrices d'alignement T^{te} et $T^{t\phi}$ disparaissent [4] puisque il n'y a plus de modèle NMF conjoint pour la source de musique dans les étapes *GEM* et *GEM-PHAT*. Dans ce scénario, durant l'étape algorithmique *NMF* (voir partie 6.2.3), les paramètres H_3^e, W_3^ϕ, H_3^ϕ, W_6, et H_6 sont mis à jour pour s'ajuster à la référence qui est déjà alignée au signal à séparer. W_6 et H_6 sont ensuite réinitialisées par des valeurs aléatoires avant que les étapes *NMPcF* et/ou *GEM* ou *GEM-PHAT* soient appliquées.

7.3.2 Combinaison algorithmique

Cette expérience utilise le même jeu de données que dans la partie 6.4, c'est-à-dire des mélanges artificiels voix/musique et des références de musique provenant de la même bande-son. Le nombre d'itérations pour les étapes *NMF* et *NMPcF* est fixé à 10 et celui des étapes *GEM* et *GEM-PHAT* qui sont connues pour nécessiter plus d'itérations est fixé à 100.

L'effet du recalage de phase de la référence de musique et de l'étape *GEM* est tout d'abord évalué. Pour ce faire, nous comparons entre la configuration avec le modèle présenté à la partie 6.4 estimé par *NMF* et *NMPcF* et la configuration qui ajoute le recalage de phase et l'étape *GEM* à cette première configuration. Les résultats sont donnés dans le Tableau 7.1. La comparaison des deuxième et troisième lignes avec la première ligne montre que l'étape *GEM* fait baisser les performances de séparation si elle est utilisée directement après l'étape *NMF* ou *NMPcF* seule. Les meilleurs résultats sont obtenus lorsque les étapes *NMF* et *NMPcF* sont toutes les deux utilisées avant l'étape *GEM*. De la même façon que l'étape *NMPcF* dépend de l'étape *NMF* (voir partie 6.4), ceci montre que le bon fonctionnement de l'étape *GEM* dépend de la combinaison de ces deux étapes. Dans ce cas, les améliorations mesurées dans nos expériences par rapport au cas non recalé en phase (équation (6.27)) sont marginales lorsque la musique est en premier plan, mais significatives lorsque la musique est en arrière-plan. Cette première constatation mène à choisir les étapes *NMF* et *NMPcF* comme techniques d'initialisation également pour l'étape *GEM-PHAT* pour cette expérience et les suivantes. Le remplacement de l'étape *GEM* par l'étape *GEM-PHAT* n'apporte cependant pas d'amélioration et dans certain cas il fait même diminuer les performances par rapport à l'initialisation (*NMF*+*NMPcF*). Une explication possible est que les valeurs de l'*annealing* (σ_f^2) ont été spécialement choisies pour l'algorithme

4. Ces matrices et le modèle (6.27) restent utilisés dans les étapes *NMF* et *NMPcF*.

	rapport voix/musique : −6 dB						rapport voix/musique : 12 dB					
	voix			musique			voix			musique		
	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
Init + NMF + NMPcF	**2,1**	2,9	8,1	9,2	11,6	17,7	6,0	8,7	24,6	0,5	2,7	3,9
Init + NMF + GEM	0,0	5,7	-3,3	1,8	-1,3	16,3	4,1	4,3	10,7	-8,7	-9,3	4,6
Init + NMPcF + GEM	-1,1	3,9	-1,0	5,3	11,6	8,4	3,2	3,8	27,9	-7,3	6,7	-6,0
Init + NMF + NMPcF + GEM	**2,2**	3,7	7,5	**9,8**	11,4	17,7	**7,6**	13,0	21,6	**2,9**	4,0	10,0
Init + NMF + NMPcF + GEM-PHAT	1,7	2,7	6,8	9,3	11,3	14,7	6,0	9,1	25,2	-1,6	2,9	1,6

Tableau 7.1 – Moyennes des performances de séparation voix/musique (dB) pour différentes combinaisons d'étapes algorithmiques dans le cas de l'utilisation d'une référence de musique recalée en phase et d'aucune référence de voix.

GEM et non pour l'algorithme *GEM-PHAT*. La présence de paramètres supplémentaires pourrait nécessiter d'autres valeurs.

Du fait de la durée d'une itération *GEM* ou *GEM-PHAT* (environ dix fois plus que *NMF* ou *NMPcF*), la pertinence de l'utilisation de ces dernières et coûteuses étapes est discutable. Au vue des améliorations marginales lorsque la musique est au premier plan, elles ne sont pas nécessaires pour ce rapport voix/musique. Inversement, lorsque la musique est en arrière-plan, l'étape *GEM* améliore le SDR de la musique de 2,4 dB. Ce résultat est également supérieur de 1,3 dB par rapport aux approches utilisant plusieurs références de musique et de voix (voir Tableau 6.3). Ce dernier point rend ces deux algorithmes intéressants dans la situation où le signal commun est faible.

7.3.3 Références synthétiques de musique

Dans cette seconde expérience, nous nous intéressons plus particulièrement à l'évaluation de l'algorithme *GEM-PHAT* en situation contrôlée de décalage de phase.

Données Le processus de création de mélange est le même que celui de la partie 6.4, à la différence que les extraits de musique sont tirés de véritables morceaux de musique et non de films. Les références de musique sont, elles, obtenues à partir de ces mêmes extraits. Deux types de déformations sont appliquées à la vérité terrain pour générer les différentes références :

— un **délai aléatoire** pour chaque trame compris entre -256 et +256 échantillons,
— un **bruit additif** de puissance −6 dB composé de parole.

Enfin le cas « oracle » est celui où la vérité terrain non déformée est utilisée comme référence. Ses résultats sont donnés à titre indicatif.

Résultats Les expériences que j'ai menées cherchent à comparer les deux étapes algorithmiques *GEM* et *GEM-PHAT* dans un scénario contrôlé de références déformées. Pour ces deux étapes, les paramètres sont initialisés comme précédemment et par les étapes *NMF* et *NMPcF* qui sont utilisées au préalable. Toutes les étapes sont itérées 10 fois. Les résultats en terme de SDR sont regroupés dans le Tableau 7.2 par type de références utilisées. Les résultats sont aussi donnés en Annexe D.3.2 regroupés par

combinaison algorithmique ainsi que les résultats à l'issue des étapes *NMF* et *NMPcF* à titre de comparaison.

On remarque que dans les deux cas où la phase de la référence a été déformée (délai aléatoire), l'étape *GEM-PHAT* donne des performances légèrement supérieures ou égales à l'étape *GEM*. En revanche lorsque la référence est uniquement bruitée, l'étape *GEM-PHAT* mène à des performances inférieures. L'étape *GEM-PHAT* semble bien modéliser la déformation de la phase sans pour autant qu'il y ait un véritable impact sur les performances en terme de SDR. De plus, l'ajout des matrices de délai $\mathbf{D}_{3,fn}$ lorsqu'il n'y pas lieu semble perturber le modèle de rang plein. Les résultats sont donc cohérents à défaut d'être conséquents.

		rapport voix/musique : −6 dB		rapport voix/musique : 12 dB	
Référence	Algorithme	voix	musique	voix	musique
Oracle	*GEM*	4,24	12,52	**7,81**	**5,38**
	GEM-PHAT	**4,49**	**13,04**	7,58	4,54
Délai aléatoire	*GEM*	0,89	6,98	7,57	4,11
	GEM-PHAT	**0,92**	**7,03**	**7,64**	**4,15**
Bruit additif	*GEM*	**4,27**	**12,52**	**7,38**	**2,63**
	GEM-PHAT	3,63	11,17	7,11	2,09
Délai aléatoire + bruit additif	*GEM*	1,15	7,15	7,05	1,81
	GEM-PHAT	**1,26**	**7,16**	**7,07**	**1,92**

Tableau 7.2 – Moyenne des SDR (dB) pour différentes déformations de la référence de musique.

7.3.4 Bandes-son dans différentes langues

Une des applications visées par ce chapitre est l'extraction de musique dans des bandes-son de films à partir de plusieurs versions en différentes langues. Dans ce cas, les différentes versions du film sont considérées comme les canaux d'un mélange regroupant toutes ces versions, par exemple les versions anglaise et française d'un épisode de série. De la même façon que pour la situation où les références de musiques proviennent d'un autre endroit dans le film (voir partie 7.3.2 [5]), on peut ici aussi utiliser les algorithmes multicanaux *GEM* ou *GEM-PHAT* après avoir recalé en phase les signaux.

Mélanges réels Les exemples de bandes-son sont fournis par notre partenaire industriel et représentent des situations réelles de post-production. Ainsi, aucune vérité terrain n'est disponible. L'évaluation d'algorithmes dans cette situation ne peut être faite de façon objective comme dans les expériences précédentes mais uniquement à l'écoute. Cela correspond à une situation réelle de post-production, où un ingénieur du son est amené à évaluer subjectivement les résultats de séparation de différents algorithmes ou de différentes paramétrisations d'un algorithme avant de conserver le meilleur résultat.

5. La même situation est aussi traité dans la partie 6.4 mais pas par un algorithme multicanal.

7.4 Conclusion

La séparation de signaux communs est un problème de séparation où le même signal est observé au sein de différents signaux. C'est le cas par exemple des enregistrements réels multicanaux ou de la musique dans plusieurs versions du même film. Cependant, ce signal peut avoir subi des déformations. Ce chapitre s'est en particulier intéressé au cas où ces déformations sont provoquées par l'utilisation d'appareils analogiques de post-production.

Ce chapitre a proposé une approche de séparation multicanale incluant un recalage de phase au cours du temps entre les canaux des sources estimées. Alors que les approches courantes se basent soit sur l'hypothèse d'un délai constant entre les canaux soit sur un alignement estimé au préalable, l'intérêt de cette approche réside dans cette estimation conjointe. On cherche ainsi réduire les erreurs commises par un recalage au préalable des mélanges alors qu'uniquement les sources communes sont concernées par le recalage.

La partie expérimentale 7.3 a exploré l'utilisation des algorithmes multicanaux *GEM-PHAT* (proposé dans ce chapitre) et *GEM* pour la séparation voix/musique guidée par une référence de musique supposée contenir un signal commun. Un premier scénario est celui où la référence de musique est issue d'une recherche de motifs au sein du même film. Un deuxième scénario plus classique est celui dit d'extraction de musique dans les bandes-son de films à partir de plusieurs versions en différentes langues. Des signaux synthétique censés représenter le problème ont aussi été utilisés pour valider l'approche. L'apport de cette nouvelle approche au vu des différents résultats est cependant assez faible. Mais il semble que différentes méthodes d'alignement temporel soient adaptées à différentes situations. En particulier, certains signaux de référence sont mieux exploités quand ils sont considérés comme des canaux du mélanges à la façon du problème de séparation de signaux communs. Par ailleurs, ces situations peuvent cohabiter dans le cadre général multicanal proposé dans le chapitre 6.

Perspectives On remarquera que l'algorithme *GEM-PHAT* est compatible avec le GEM de la partie 6.2. En effet, ces deux algorithmes sont des extensions du cadre général proposé dans [136]. Les modifications apportées par l'algorithme *GEM-PHAT* concernent les paramètres spatiaux, alors que le GEM précédemment présenté apporte des modifications aux modèles spectraux et étend à M mélanges la formulation du problème de séparation. De plus, la combinaison d'une référence de musique recalée en phase et d'une référence de parole (6.26) n'a pas été testée et pourrait améliorer l'initialisation du modèle spectral. L'utilité du modèle spectral restant à prouver pour l'étape multicanale, il pourrait ensuite être réduit à un simple scalaire v_{fn} pour cette étape.

Troisième partie

Insertion industrielle et perspectives

Chapitre 8

Insertion industrielle des travaux de thèse

[Confidentiel]

Chapitre 9

Conclusion et perspectives scientifiques

9.1 Conclusion

L'approche abordée dans cette thèse, dénommée *SPotted REference based Separation* ou SPORES, vise à exploiter la redondance des contenus audiovisuels pour la séparation de sources. L'hypothèse de travail consiste à détecter des motifs sonores répétés et à les considérer comme signaux de référence pour la séparation de sources.

La première partie de ce manuscrit a permis de formaliser les problèmes de recherche de motifs et de séparation de sources et de dresser leurs états de l'art respectifs avec notamment un focus sur les techniques de séparation de sources guidée. Cette phase de travail à également permis d'identifier les verrous scientifiques à faire sauter pour mettre en pratique le concept SPORES. D'une part, les techniques de détection de motifs ne sont pas adaptées pour traiter des mélanges (requêtes ou motifs recherchés). D'autre part, les techniques de séparation guidée sont généralement développées pour un unique scénario de séparation. D'autres approches exploitant les répétitions des signaux ont été comparées au concept SPORES dans sa globalité. Les approches dites informées sont adaptées à la situation où les sources originelles sont observés seules avec comme but leur transmission, ce qui n'est pas le cas de SPORES. Les approches de type REPET se basent sur les ressemblances entre trames alors que SPORES tire partie de l'ensemble du motif.

La deuxième partie regroupait les différentes contributions scientifiques pour répondre à une partie des verrous identifiés, avec notamment des aspects algorithmiques, des propositions de modèles et des nouveaux scénarios d'utilisation. Une première étude sur le choix des distances de comparaison dans le domaine STFT a été menée avec comme but principal de rendre les techniques de détection de motifs plus robustes à la présence d'autres sources. Les distances l_p pour $p \leq 0{,}2$ ont notamment montré une meilleure précision que la distance cosinus sur les premières réponses de la détection. Un modèle général de déformation pour les références a ensuite été proposé et validé sur différentes tâches de séparation. Plusieurs facteurs déterminants ont été identifiés,

notamment l'utilisation des références pour l'initialisation des modèles, l'importance de détenir au moins $J-1$ références pour estimer J sources et enfin la qualité des références. Dans le cas particulier de signaux communs, c'est-à-dire lorsque la source de référence a une forme d'onde proche de celle de la source à estimer, un autre modèle a été proposé. Il permet de prendre en compte des déformations temporelles plus fines que le modèle général. Ce modèle a cependant peu d'impact sur la qualité de la séparation. Les expériences ont montré que l'utilisation de références pour une source donnée améliore la qualité audio globale des sources estimées (de 9 à 15 dB).

La troisième et dernière partie a décrit l'étude de l'insertion effective des résultats de ces recherches dans le flux de travail du Studio Maia. Elle met principalement en évidence la complémentarité des outils issus des travaux de cette thèse avec les outils déjà existants. De multiples retombés à plus long terme de ces travaux ont aussi été mises en avant que ce soit pour la création de nouveaux outils pour l'entreprise partenaire ou pour de futures applications.

9.2 Perspectives scientifiques

J'ai évoqué tout au long du manuscrit des perspectives pour les différents travaux effectués durant cette thèse. Je donne dans cette partie une description détaillée de pistes dans le prolongement de ce travail.

Distances l_p La comparaison d'une trame à une autre est utilisée par de nombreuses approches comme opération élémentaire. Les propriétés de cette opération élémentaire impactent généralement les propriétés de l'approche qui les utilise. Par exemple un moyen de comparaison robuste à la présence d'autres sources rendrait possible le traitement de mélanges pour d'autres tâches. C'est le cas des distances l_p étudiées dans le chapitre 5 qui peuvent apporter une plus-value aux approches utilisant des matrices de similarité qui ont déjà été décrites dans ce manuscrit :

- la découverte de motifs pour le calcul de la DTW (voir partie 2.4.3), les motifs pouvant alors apparaitre en présence d'autres sources [1],
- l'extraction du fond musical aveugle de type *REPET-SIM* [145] (voir partie 4.2) pour identifier les trames similaires (voir partie 5.3 pour une discussion détaillée),
- la séparation de sources guidée par référence pour l'initialisation des matrices d'alignement (DTW et pondération) (voir partie 6.4), la présence d'autres sources dans le mélange ou la référence pouvant être pris en compte par des distances tolérantes à leurs présences,
- la technique non itérative de séparation de sources guidée par référence (*RbWF*) également pour l'initialisation des matrices d'alignement (voir partie 6.4).

Ces perspectives sont aussi valables pour des techniques de comparaison tolérantes aux déformations comme pour la découverte de motifs [123].

1. Une limitation majeure du formalisme ARGOS [88] suivi dans [34, 123] est alors la non prise en compte de la superposition de motifs. Il ne peut donc pas en l'état traiter correctement des mélanges.

Séparation guidée Dans ce manuscrit, les expériences de séparation guidée ont exploré de nombreuses facettes de ce problème. Il reste, toutefois, des pistes de travail, par exemple en ce qui concerne la modélisation des déformations avec notamment l'utilisation de matrices T_f variant au cours du temps ou encore la prise en compte de déformations non linéaires. Concernant la séparation de signaux communs, modéliser l'évolution temporelle du décalage de phase est une hypothèse de travail. Ce suivi pourrait par exemple être opéré par un HMM comme dans l'approche dénommée DOA-HMM [90] où l'évolution temporelle de l'angle d'arrivé est modélisée par un HMM, chaque zone angulaire étant représentée par un état du HMM.

Approche SPORES non supervisée L'approche SPORES telle qu'elle a été décrite et utilisée par Maia, est une approche de séparation qui se focalise sur le traitement des segments audio pour lesquels des signaux de référence sont fournis par détection de motifs. Un utilisateur intervient au cours du processus sur le choix des références à utiliser. Une perspective est l'automatisation de l'ensemble du processus pour l'ensemble des segments audio, typiquement un film entier ou un morceau de musique. La phase de recherche de motifs devient alors une phase de découverte de motifs, et plus aucune requête de recherche n'est spécifiée. Les occurrences de chaque motif découvert servent ensuite de références les unes pour les autres pour la phase de séparation guidée par référence.

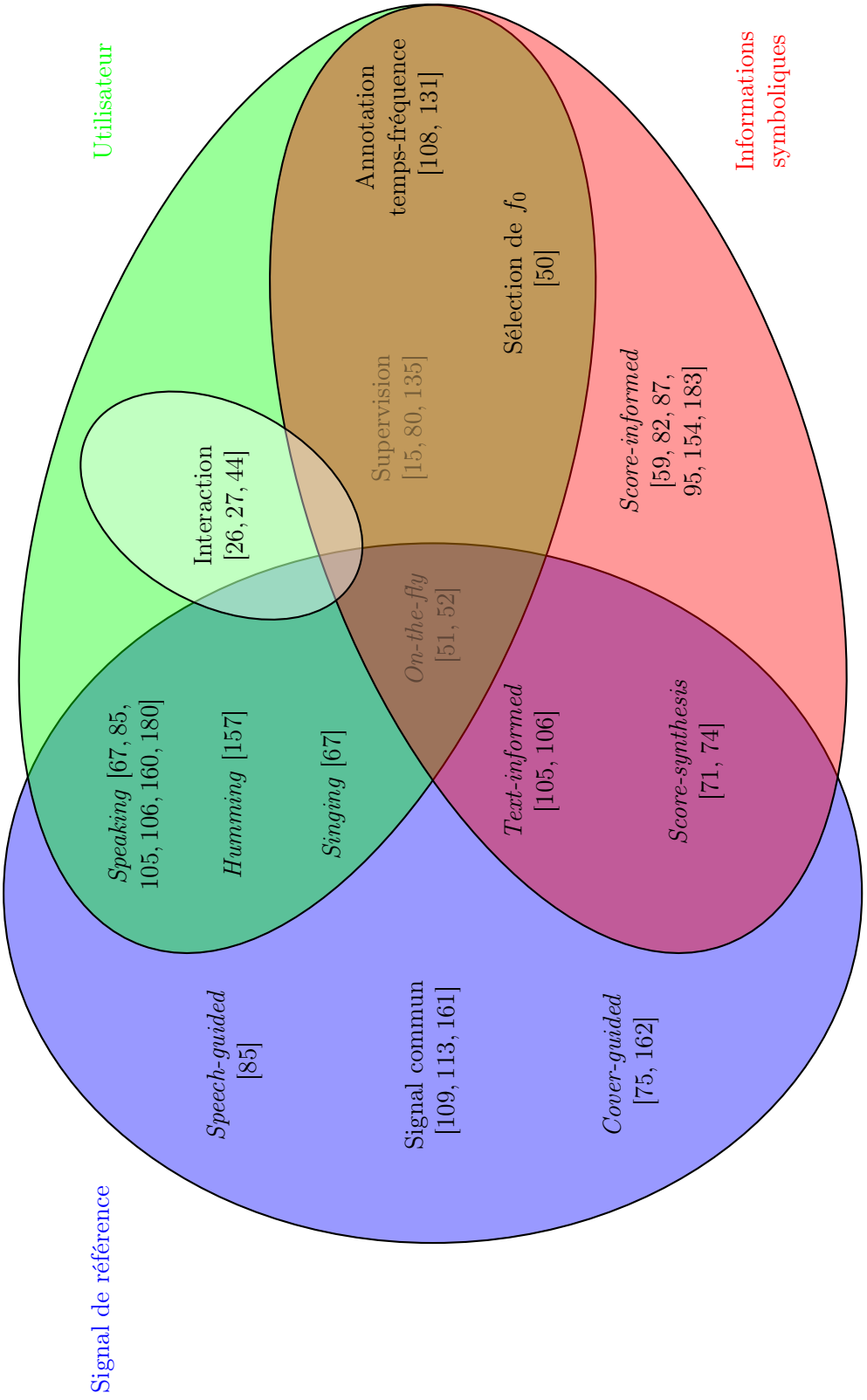
Le problème majeur de cette approche globale et non supervisée est le fait de ne pas avoir de signaux de référence pour toutes les sources et tous les segments audio. Une première possibilité est de séparer dans un segment donné uniquement les sources ayant une référence, plus une dernière source regroupant les éléments sonores non redondants (par exemple la voix chantée dans un morceau de musique). Il serait ensuite nécessaire d'agglomérer à posteriori les différents segments séparés pour former les sources séparées globales.

On peut aussi imaginer que les segments ayant $J-1$ références soient séparés dans un premier temps, puis les segments avec une référence de moins et ainsi de suite jusqu'à traiter les segments sans références. On pourrait ainsi établir à partir des sources séparées un modèle ou des signaux de référence à priori pour chaque source après chaque phase de séparation. Ils pourraient ensuite être utilisés comme substituts en l'absence de référence pour la source correspondante. Cette piste de travail repose sur l'hypothèse que $J-1$ sources sont identifiables par leur redondance.

Annexes

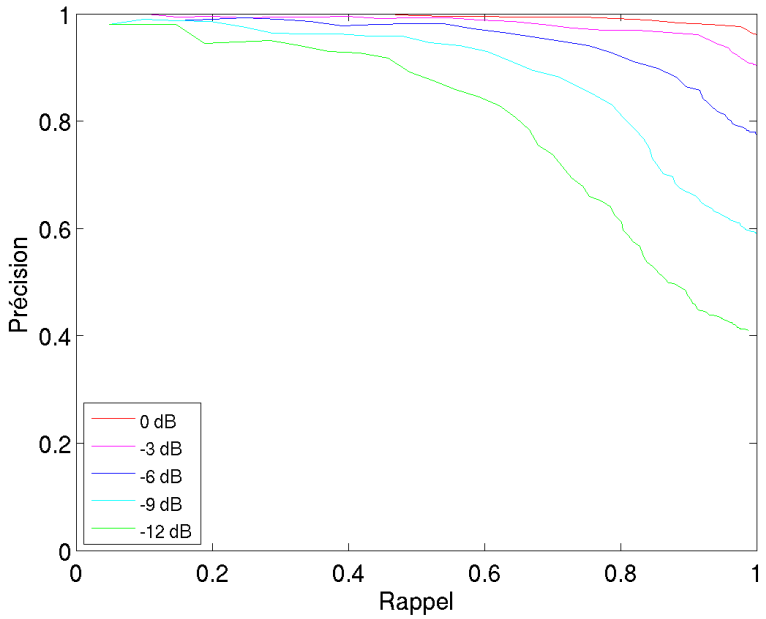
Annexe A

Classification des techniques de séparation de sources fortement guidée

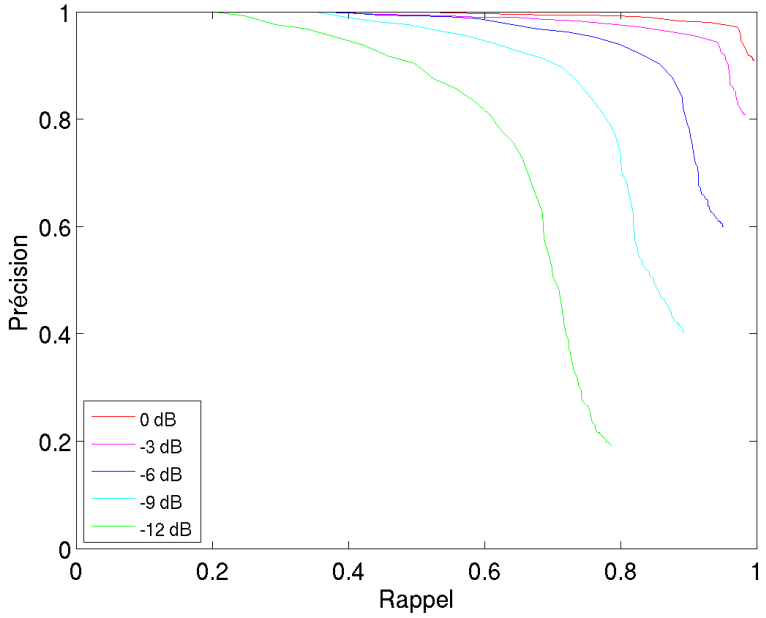

Utilisateur
Informations symboliques
Signal de référence
Annotation temps-fréquence [108, 131]
Sélection de f_0 [50]
Supervision [15, 80, 135]
Interaction [26, 27, 44]
On-the-fly [51, 52]
Score-informed [59, 82, 87, 95, 154, 183]
Text-informed [105, 106]
Score-synthesis [71, 74]
Speaking [67, 85, 105, 106, 160, 180]
Humming [157]
Singing [67]
Speech-guided [85]
Signal commun [109, 113, 161]
Cover-guided [75, 162]

Annexe B

Courbes précision-rappel



(a) Distance cosinus.



(b) Distance l_p avec $p = 0{,}1$

Figure B.1 – Courbes précision-rappel des distances cosinus et l_p avec $p = 0{,}1$ pour différents niveaux (de 0 à -12 dB) des motifs dans les mélanges de recherche pour la tâche de détection de motifs musicaux en présence de parole (niveau des motifs dans la requête : -12 dB).

Annexe C

Calculs détaillés

C.1 Calcul détaillé de l'espérance de la log-vraisemblance des données complètes (6.18)

$$
\begin{aligned}
Q(\theta,\theta^c) &\triangleq \mathbb{E}_{\mathbf{Z}|\theta^c}\left[\log p(\mathbf{Z}|\theta)\right] = \sum_m \lambda^m \mathbb{E}_{\mathbf{Z}|\theta^c}[\log p(\mathbf{X}^m,\mathbf{S}^m|\theta)] \qquad \text{(C.1)}\\
&= \sum_m \lambda^m \mathbb{E}_{\mathbf{X}^m,\mathbf{S}^m|\theta^c}[\log p(\mathbf{X}^m|\mathbf{S}^m)] + \sum_m \lambda^m \mathbb{E}_{\mathbf{S}^m|\theta^c}[\log p(\mathbf{S}^m|\theta)]\\
&= \sum_{m,f,n} \lambda^m \mathbb{E}_{\mathbf{x}^m_{fn},\mathbf{s}^m_{fn}|\theta^c}\left[\log \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}^m_{fn}|\mathbf{A}^m_{fn}\mathbf{s}^m_{fn},\mathbf{\Sigma}_{\mathbf{b}^m_{fn}}\right)\right]\\
&\quad + \sum_{m,j\in\mathcal{J}^m,f,n} \lambda^m \sum_{r=1}^{R_j} \mathbb{E}_{s_{jr,fn}|\theta^c}\left[\log \mathcal{N}_{\mathbb{C}}\left(s_{jr,fn}|0,v_{j,fn}\right)\right]\\
&\overset{c}{=} -\sum_{mfn} \frac{\lambda^m}{\sigma_f^2}\mathrm{tr}\left[\mathbf{R}_{\mathbf{x}^m_{fn}} - \mathbf{A}^m_{fn}\mathbf{R}_{\mathbf{x}\mathbf{s}^m_{fn}}{}^H - \mathbf{R}_{\mathbf{x}\mathbf{s}^m_{fn}}\mathbf{A}^{m\,H}_{fn} + \mathbf{A}^m_{fn}\mathbf{R}_{\mathbf{s}^m_{fn}}\mathbf{A}^{m\,H}_{fn}\right]\\
&\quad - \sum_{m,j\in\mathcal{J}^m,fn} \lambda^m R_j d_{IS}(\xi_{j,fn}|v_{j,fn})
\end{aligned}
$$

C.2 Calcul détaillé de la mise à jour des paramètres spatiaux pour l'étape M de l'algorithme GEM-PHAT (7.9).

On dénote $\mathcal{J}\backslash j$ l'ensemble des sources dont on exclut j. La log-vraisemblance s'écrit en fonction du paramètre spatial $\mathbf{\Lambda}_{j,f}$ comme

$$
\begin{aligned}
Q(\theta,\theta^c) &= -\sum_{fn}\frac{1}{\sigma_f^2}\mathrm{tr}\left[\cancel{\widehat{\mathbf{R}}_{\mathbf{x}_{fn}}} - \mathbf{A}_{fn}\widehat{\mathbf{R}}^H_{\mathbf{xs}_{fn}} - \cancel{\widehat{\mathbf{R}}_{\mathbf{xs}_{fn}}\mathbf{A}_{fn}^H} + \mathbf{A}_{fn}\widehat{\mathbf{R}}_{\mathbf{s}_{fn}}\mathbf{A}_{fn}{}^H\right] + \mathrm{cst} \qquad \text{(C.2)}\\
&= -\sum_{\bar{\jmath}\in\mathcal{J},fn}\frac{1}{\sigma_f^2}\mathrm{tr}\left(\mathbf{D}_{\bar{\jmath},fn}\mathbf{\Lambda}_{\bar{\jmath},f}\widehat{\mathbf{R}}^H_{\mathbf{xs}_{\bar{\jmath},fn}}\right) + \sum_{\bar{\jmath}\in\mathcal{J},\tilde{\jmath}\in\mathcal{J},fn}\frac{1}{\sigma_f^2}\mathrm{tr}\left(\mathbf{D}_{\bar{\jmath},f}\mathbf{\Lambda}_{\bar{\jmath},f}\widehat{\mathbf{R}}_{\mathbf{s}_{\bar{\jmath}\tilde{\jmath},fn}}(\mathbf{D}_{\tilde{\jmath},f}\mathbf{\Lambda}_{\tilde{\jmath},f})^H\right) + \mathrm{cst}\\
&= -\sum_{fn}\frac{1}{\sigma_f^2}\mathrm{tr}\left(\mathbf{D}_{j,fn}\mathbf{\Lambda}_{j,f}\widehat{\mathbf{R}}^H_{\mathbf{xs}_{j,fn}}\right) - \cancel{\sum_{\bar{\jmath}\in\mathcal{J}\backslash j,fn}\frac{1}{\sigma_f^2}\mathrm{tr}\left(\mathbf{D}_{\bar{\jmath},fn}\mathbf{\Lambda}_{\bar{\jmath},f}\widehat{\mathbf{R}}^H_{\mathbf{xs}_{\bar{\jmath},fn}}\right)} + \mathrm{cst}\\
&\quad + \sum_{\tilde{\jmath}\in\mathcal{J},fn}\frac{1}{\sigma_f^2}\mathrm{tr}\left(\mathbf{D}_{j,f}\mathbf{\Lambda}_{j,f}\widehat{\mathbf{R}}_{\mathbf{s}_{j\tilde{\jmath},fn}}(\mathbf{D}_{\tilde{\jmath},f}\mathbf{\Lambda}_{\tilde{\jmath},f})^H\right) + \cancel{\sum_{\bar{\jmath}\in\mathcal{J}\backslash j,\tilde{\jmath}\in\mathcal{J},fn}\frac{1}{\sigma_f^2}\mathrm{tr}\left(\mathbf{D}_{\bar{\jmath},f}\mathbf{\Lambda}_{\bar{\jmath},f}\widehat{\mathbf{R}}_{\mathbf{s}_{\bar{\jmath}\tilde{\jmath},fn}}(\mathbf{D}_{\tilde{\jmath},f}\mathbf{\Lambda}_{\tilde{\jmath},f})^H\right)}\\
&= -\sum_{fn}\frac{1}{\sigma_f^2}\mathrm{tr}\left(\mathbf{D}_{j,fn}\mathbf{\Lambda}_{j,f}\widehat{\mathbf{R}}^H_{\mathbf{xs}_{j,fn}}\right) + \sum_{fn}\frac{1}{\sigma_f^2}\mathrm{tr}\left(\mathbf{D}_{j,f}\mathbf{\Lambda}_{j,f}\widehat{\mathbf{R}}_{\mathbf{s}_{jj,fn}}(\mathbf{D}_{j,f}\mathbf{\Lambda}_{j,f})^H\right)\\
&\quad + \sum_{\tilde{\jmath}\in\mathcal{J}\backslash j,fn}\frac{1}{\sigma_f^2}\mathrm{tr}\left(\mathbf{D}_{j,f}\mathbf{\Lambda}_{j,f}\widehat{\mathbf{R}}_{\mathbf{s}_{j\tilde{\jmath},fn}}(\mathbf{D}_{\tilde{\jmath},f}\mathbf{\Lambda}_{\tilde{\jmath},f})^H\right) + \mathrm{cst}
\end{aligned}
$$

où « cst » dénote un terme indépendant de $\mathbf{\Lambda}_{j,f}$. $\mathbf{\Lambda}^H_{j,f}$ ne dépend pas de $\mathbf{\Lambda}_{j,f}$ lorsqu'il s'agit de calculer la dérivée complexe. Après avoir retiré la somme sur les fréquences f, la dérivée de Q par rapport à $\mathbf{\Lambda}_{j,f}$ s'écrit alors

$$
-\frac{1}{\sigma_f^2}\sum_n\left[\mathbf{D}^{\mathscr{T}}_{\bar{\jmath},fn}(\widehat{\mathbf{R}}_{\mathbf{xs}_{j,fn}})^\star\right] + \frac{1}{\sigma_f^2}\sum_n\left[\cancel{\mathbf{D}^{\mathscr{T}}_{\bar{\jmath},fn}\mathbf{D}^\star_{j,fn}}\mathbf{\Lambda}^\star_{j,f}\widehat{\mathbf{R}}_{\mathbf{s}_{jj,fn}}\right] + \frac{1}{\sigma_f^2}\sum_{\bar{\jmath}\in\mathcal{J}\backslash j,n}\left[\mathbf{D}^{\mathscr{T}}_{\bar{\jmath},fn}\mathbf{D}^\star_{\tilde{\jmath},fn}\mathbf{\Lambda}^\star_{\tilde{\jmath},f}\widehat{\mathbf{R}}_{\mathbf{s}_{\tilde{\jmath}j,fn}}\right]. \qquad \text{(C.3)}
$$

On cherche à annuler cette dérivée ($\frac{\partial Q(\theta,\theta^c)}{\partial \mathbf{\Lambda}_{j,f}} = \mathbf{0}$) :

$$
\sum_n\left[\mathbf{\Lambda}^\star_{j,f}\widehat{\mathbf{R}}_{\mathbf{s}_{jj,fn}}\right] = \sum_n\left[\mathbf{D}_{j,fn}(\widehat{\mathbf{R}}_{\mathbf{xs}_{j,fn}})^\star\right] - \sum_{\tilde{\jmath}\in\mathcal{J}\backslash j,n}\left[\mathbf{D}_{j,fn}\mathbf{D}^\star_{\tilde{\jmath},fn}\mathbf{\Lambda}^\star_{\tilde{\jmath},f}\widehat{\mathbf{R}}_{\mathbf{s}_{\tilde{\jmath}j,fn}}\right] \qquad \text{(C.4)}
$$

$$
\Rightarrow \mathbf{\Lambda}_{j,f} = \left[\sum_n \mathbf{D}^H_{j,fn}\left\{\widehat{\mathbf{R}}_{\mathbf{xs}_{j,fn}} - \sum_{\tilde{\jmath}\in\mathcal{J}\backslash j}\mathbf{D}_{\tilde{\jmath},fn}\mathbf{\Lambda}_{\tilde{\jmath},f}\widehat{\mathbf{R}}_{\mathbf{s}_{\tilde{\jmath}j,fn}}\right\}\right]\left[\sum_n \widehat{\mathbf{R}}_{\mathbf{s}_{jj,fn}}\right]^{-1}, \qquad \text{(C.5)}
$$

Certaines simplifications proviennent du fait que :
— $\forall j$, $\mathbf{D}_{j,fn}\mathbf{D}^\star_{j,fn} = \mathbf{I}$, $\mathbf{D}^T_{j,fn} = \mathbf{D}_{j,fn}$ et $\mathbf{D}^H_{j,fn} = \mathbf{D}^\star_{j,fn}$
— $\widehat{\mathbf{R}}_{\mathbf{s}_{fn}} = (\widehat{\mathbf{R}}_{\mathbf{s}_{fn}})^H$
et des propriétés des matrices complexes suivantes [140] :
— $\frac{\partial \mathrm{tr}(\mathbf{AXB})}{\partial \mathbf{X}} = \mathbf{A}^T\mathbf{B}^T$

— $\frac{\partial \mathrm{tr}(\mathbf{X}^H)}{\partial \mathbf{X}} = \mathbf{0}$
— $\mathbf{A}^H = (\mathbf{A}^\star)^T$
— $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$, $(\mathbf{AB})^\star = \mathbf{A}^\star\mathbf{B}^\star$ et $(\mathbf{AB})^H = \mathbf{B}^H\mathbf{A}^H$

Annexe D

Tableaux

D.1 Tableaux complémentaires pour l'apprentissage de p pour un instrument dans un morceau de musique

	-12 **dB**	-9 **dB**	-6 **dB**	-3 **dB**	0 **dB**	$+\infty$
-12 **dB**	0,32	0,32	0,31	0,30	0,30	0,28
-9 **dB**	0,32	0,32	0,32	0,31	0,30	0,28
-6 **dB**	0,31	0,32	0,32	0,32	0,31	0,27
-3 **dB**	0,31	0,31	0,32	0,32	0,32	0,27
0 **dB**	0,30	0,31	0,31	0,32	0,33	0,27

(a) Apprentissage pour la batterie.

	-12 **dB**	-9 **dB**	-6 **dB**	-3 **dB**	0 **dB**	$+\infty$
-12 **dB**	0,39	0,38	0,37	0,37	0,36	0,39
-9 **dB**	0,39	0,39	0,38	0,37	0,37	0,38
-6 **dB**	0,38	0,39	0,39	0,38	0,37	0,38
-3 **dB**	0,38	0,38	0,39	0,39	0,38	0,37
0 **dB**	0,38	0,38	0,38	0,39	0,39	0,37

(b) Apprentissage pour la guitare.

Tableau D.1 – Valeur moyenne de p entre deux occurrences d'un instrument de musique mélangé à deux niveaux différents dans des portions différentes du morceau avec tous les instruments.

D.2 Intervalles de confiance pour la comparaison des courbes précision-rappel

Rappel	0.1	0.15	0.19	0.24	0.29	0.33	0.37	0.41
$p = 0.1$	92	96	99.9	99.7	93	90	81	65
$p = 0.2$	92	96	99.9	99.4	86	77	65	58

Tableau D.2 – Probabilités l'hypothèse non-nulle, c'est-à-dire de non-égalité entre les valeurs de précisions pour la distance cosinus et les distances $l_{0.1}$ et $l_{0.2}$ pour des valeurs de rappel inférieures à 0.4. Les valeurs sont données pour l'expérience de détection avec des niveaux de la requête et des motifs recherchés de -12dB.

D.3 Tableaux complets pour la séparation voix/musique avec une référence de musique

D.3.1 Combinaison algorithmique

		rapport voix/musique : −6 dB						rapport voix/musique : 12 dB					
		voix			musique			voix			musique		
		SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
Sans recalage de la phase	*Init + NMPcF*	-0,6	2,3	-2,1	5,9	11,7	8,3	3,6	4,9	29,2	-6,8	6,8	-5,5
	Init + NMF + Plain-NMF	1,8	1,0	8,9	9,2	13,1	12,4	5,1	7,1	23,8	-3,9	3,1	-1,5
	Init + NMF + NMPcF	**2,1**	2,9	8,1	**9,2**	11,6	17,7	6,0	8,7	24,6	**0,5**	2,7	3,9
Recalage de la phase	*Init + NMF + GEM*	0,0	5,7	-3,3	1,8	-1,3	16,3	4,1	4,3	10,7	-8,7	-9,3	4,6
	Init + NMPcF + GEM	-1,1	3,9	-1,0	5,3	11,6	8,4	3,2	3,8	27,9	-7,3	6,7	-6,0
	Init + NMF + NMPcF + GEM	**2,2**	3,7	7,5	**9,8**	11,4	17,7	**7,6**	13,0	21,6	**2,9**	4,0	10,0
	Init + NMF + NMPcF + GEM-PHAT	1,7	2,7	6,8	9,3	11,3	14,7	6,0	9,1	25,2	-1,6	2,9	1,6

Tableau D.3 – Moyennes des performances de séparation voix/musique (dB) pour différentes combinaisons d'étapes algorithmiques dans le cas de l'utilisation d'une référence de musique et d'aucune référence de voix. Les meilleurs SDR sont indiqués en gras pour chaque colonne par partie (délimitée par des doubles lignes).

D.3.2 Référence de musique déformée synthétiquement

	rapport voix/musique : -6 dB		rapport voix/musique : 12 dB	
Référence	voix	musique	voix	musique
Oracle	4,65	13,30	7,44	4,19
Délai aléatoire	-0,03	5,65	7,45	3,57
Bruit additif	3,63	11,22	6,97	1,71
Délai aléatoire + bruit additif	1,10	6,81	6,79	1,10

Tableau D.4 – Moyenne des SDRs (dB) pour différentes déformations de la référence de musique pour l'algorithme *NMF+NMPcF* seul.

	rapport voix/musique : -6 dB		rapport voix/musique : 12 dB	
Référence	voix	musique	voix	musique
Oracle	4,24	12,52	7,81	5,38
Délai aléatoire	0,89	6,98	7,57	4,11
Bruit additif	4,27	12,52	7,38	2,63
Délai aléatoire + bruit additif	1,15	7,15	7,05	1,81

Tableau D.5 – Moyenne des SDRs (dB) pour différentes déformations de la référence de musique pour l'algorithme *GEM*.

	rapport voix/musique : -6 dB		rapport voix/musique : 12 dB	
Référence	voix	musique	voix	musique
Oracle	4,49	13,04	7,58	4,54
Délai aléatoire	0,92	7,03	7,64	4,15
Bruit additif	3,63	11,17	7,11	2,09
Délai aléatoire + bruit additif	1,26	7,16	7,07	1,92

Tableau D.6 – Moyenne des SDRs (dB) pour différentes déformations de la référence de musique pour l'algorithme *GEM-PHAT*.

Bibliographie

[1] *IEEE Signal Processing Magazine*, volume 31(3). May 2014.

[2] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley. Audio inpainting. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3) :922–932, March 2012.

[3] J. Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Audio, Speech and Language Processing*, 25(3) :235–238, Jun 1977.

[4] S. Arberet, R. Gribonval, and F. Bimbot. A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture. In *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *Lecture Notes in Computer Science*, pages 536–543. Springer Berlin Heidelberg, 2006.

[5] S. Arberet, R. Gribonval, and F. Bimbot. A robust method to count and locate audio sources in a stereophonic linear anechoic mixture. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 745–748, Honolulu, HI, United States, April 2007.

[6] S. Arberet, A. Ozerov, F. Bimbot, and R. Gribonval. A tractable framework for estimating and combining spectral source models for audio source separation. *Signal Processing*, 92(8) :1886 – 1901, 2012. Latent Variable Analysis and Signal Separation.

[7] Archivage. In *Dictionnaire Larousse en ligne*, 2015.

[8] C. Avendano. Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 55–58, New Paltz, NY, United States, Oct 2003.

[9] C. Avendano and J.-M. Jot. Frequency domain techniques for stereo to multichannel upmix. In *Proc. 22nd Audio Engineering Society (AES) Convention*, Jun 2002.

[10] J. Barbedo and G. Tzanetakis. Musical instrument classification using individual partials. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1) :111–122, Jan 2011.

[11] M. Bartsch and G. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1) :96–104, Feb 2005.

[12] M. Bay and J. W. Beauchamp. Multiple-timbre fundamental frequency tracking using an instrument spectrum library. *The Journal of the Acoustical Society of America*, 132(3) :1886–1886, 2012.

[13] R. Bellman. *Dynamic Programming*. Dover Publications, 1957.

[14] L. Benaroya, F. Bimbot, and R. Gribonval. Audio Source Separation With a Single Sensor. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1) :191–199, Jan. 2006.

[15] L. Benaroya, L. Donagh, F. Bimbot, and R. Gribonval. Non negative sparse representation for wiener based source separation with a single sensor. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages VI–613–16 vol.6, April 2003.

[16] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription : challenges and future directions. *Journal of Intelligent Information Systems*, 41(3) :407–434, 2013.

[17] Y. Benezeth, G. Bachman, G. Le-Jan, N. Souviraà-Labastie, and F. Bimbot. BL-Database : A french audiovisual database for speech driven lip animation systems. Technical report RR-7711, INRIA, August 2011.

[18] O. Bermond and J.-F. Cardoso. Méthodes de séparation de sources dans le cas sous-déterminé. In *Proc. of Groupe d'Etudes du Traitement du Signal et des Images (GRETSI)*, pages 749–752, Vannes, France, 1999.

[19] S.-A. Berrani, M. H. Boukadida, and P. Gros. Constraint satisfaction programming for video summarization. In *IEEE International Symposium on Multimedia*, pages 195–202, Anaheim, CA, United States, Dec. 2013. IEEE.

[20] N. Bertin. *Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique.* Thèses, Télécom ParisTech, Oct. 2009.

[21] N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3) :538–549, March 2010.

[22] J. Bitzer, K. Simmer, and K.-D. Kammeyer. Theoretical noise reduction limits of the generalized sidelobe canceller (gsc) for speech enhancement. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 2965–2968 vol.5, Phoenix, AZ, United States, 1999.

[23] J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 559–564, Porto, Portugal, Oct 2012.

[24] M. Brandstein and H. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 375–378 vol.1, Munich, Germany, Apr 1997.

[25] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers, and L. Terentiev. Spatial Audio Object Coding (SAOC) - The Upcoming MPEG Standard on Parametric Object Based Audio Coding. In *Proc. 124th Audio Engineering Society (AES) Convention*, Amsterdam, Netherlands, May 2008.

[26] N. J. Bryan and G. J. Mysore. Interactive refinement of supervised and semi-supervised sound source separation estimates. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 883–887, Vancouver, Canada, May 2013.

[27] N. J. Bryan, G. J. Mysore, and G. Wang. ISSE : an interactive source separation editor. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pages 257–266, Toronto, Canada, 2014.

[28] J. Burred. Genetic motif discovery applied to audio analysis. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 361–364, Kyoto, Japan, March 2012.

[29] J. Burred and P. Leveau. Geometric multichannel common signal separation with application to music and effects extraction from film soundtracks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 201–204, Prague, Czech Republic, May 2011.

[30] R. Cabral Farias, J. Cohen, C. Jutten, and P. Comon. Joint decompositions with flexible couplings. In E. Vincent, A. Yeredor, Z. Koldovsky, and P. Tichavsky, editors, *Latent Variable Analysis and Signal Separation*, volume 9237 of *Lecture Notes in Computer Science*, pages 119–126. Springer International Publishing, 2015.

[31] V. D. Calhoun, T. Adali, K. A. Kiehl, R. Astur, J. J. Pekar, and G. D. Pearlson. A method for multitask fmri data fusion applied to schizophrenia. *Hum. Brain Mapp*, pages 598–610, 2006.

[32] J. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4) :112–114, April 1997.

[33] J. Cardoso. Multidimensional independent component analysis. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 1941–1944 vol.4, Seattle, WA, United States, May 1998.

[34] L. Catanese, N. Souviraà-Labastie, B. Qu, S. Campion, G. Gravier, E. Vincent, and F. Bimbot. MODIS : an audio motif discovery software. In *Show & Tell - Interspeech 2013*, pages 2675–2677, Lyon, France, August 2013.

[35] B. Cheng, C. Ritz, and I. Burnett. Encoding independent sources in spatially squeezed surround audio coding. In *Advances in Multimedia Information Processing - PCM 2007*, volume 4810 of *Lecture Notes in Computer Science*, pages 804–813. Springer Berlin Heidelberg, 2007.

[36] A. Cichocki, R. Zdunek, and S.-i. Amari. Csiszár's divergences for non-negative matrix factorization : Family of new algorithms. In *Independent Component Ana-*

lysis and Blind Signal Separation, volume 3889 of *Lecture Notes in Computer Science*, pages 32–39. Springer Berlin Heidelberg, 2006.

[37] P. Comon and C. Jutten. *Handbook of Blind Source Separation : Independent component analysis and applications*. Academic press, 2010.

[38] C. Cotton and D. Ellis. Audio fingerprinting to identify multiple videos of an event. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2386–2389, Dallas, TX, United States, March 2010.

[39] R. Crochiere. A weighted overlap-add method of short-time fourier analysis/synthesis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(1) :99–102, Feb 1980.

[40] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society, series B*, 39(1) :1–38, 1977.

[41] O. Dikmen and A. T. Cemgil. Unsupervised single-channel source separation using bayesian nmf. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 93–96, New Paltz, NY, United States, 2009. IEEE.

[42] J. Dmochowski, J. Benesty, and S. Affes. A generalized steered response power method for computationally viable source localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8) :2510–2526, Nov 2007.

[43] N. Q. K. Duong. *Modélisation gaussienne de rang plein des mélanges audio convolutifs appliquée à la séparation de sources*. Thèses, Université Rennes 1, 2011.

[44] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot. An interactive audio source separation framework based on non-negative matrix factorization. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1567–1571, Florence, Italy, May 2014.

[45] N. Q. K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7) :1830–1840, Jul. 2010.

[46] E. Dupraz and G. Richard. Robust frequency-based audio fingerprinting. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 281–284, Dallas, TX, United States, March 2010.

[47] J.-L. Durrieu. *Automatic transcription and separation of the main melody in polyphonic music signals*. Thèses, Ecole nationale supérieure des telecommunications-ENST, 2010.

[48] J.-L. Durrieu, A. Ozerov, C. Févotte, and G. Richard. Main instrument separation from stereophonic audio signals using a source/filter model. In *Proc. 17th European Signal Processing Conference (EUSIPCO)*, pages 15–20, Glasgow, United Kingdom, Aug. 2009.

[49] J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3) :564–575, Mar. 2010.

[50] J.-L. Durrieu and J.-P. Thiran. Musical audio source separation based on user-selected F0 track. In *Proc. 10th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 438–445, Tel-Aviv, Israel, Mar. 2012.

[51] D. El Badawy, N. Q. K. Duong, and A. Ozerov. On-the-fly audio source separation. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2014)*, pages 1–6, Reims, France, Sept. 2014.

[52] D. El Badawy, A. Ozerov, and N. Q. K. Duong. Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 256–260, Brisbane, Queensland, Australia, Apr. 2015.

[53] D. P. W. Ellis. Classifying music audio with timbral and chroma features abstract, 2007.

[54] D. P. W. Ellis. Dynamic time warping in Matlab, 2003.

[55] D. P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005.

[56] V. Emiya. *Transcription automatique de la musique de piano.* Thèses, Télécom ParisTech, Oct. 2008.

[57] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. The PEASS Toolkit - Perceptual Evaluation methods for Audio Source Separation. 9th Int. Conf. on Latent Variable Analysis and Signal Separation, Sept. 2010.

[58] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7) :2046–2057, Sept. 2011.

[59] S. Ewert and M. Müller. Score-informed voice separation for piano recordings. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 245–250, Miami, FL, United States, October 2011.

[60] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley. Score-informed source separation for musical audio recordings : An overview. *IEEE Signal Processing Magazine*, 31(3) :116–124, May 2014.

[61] C. Faller, A. Favrot, Y.-W. Jung, and H.-O. Oh. Enhancing stereo audio with remix capability. In *Proc. 129th Audio Engineering Society (AES) Convention*, Nov 2010.

[62] G. Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, 1960.

[63] S. Fenet. *Empreintes audio et stratégies d'indexation associées pour l'identification audio à grande échelle.* Thèses, Télécom ParisTech, 2013.

[64] S. Fenet, G. Richard, and Y. Grenier. A scalable audio fingerprint method with robustness to pitch-shifting. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 121–126, Miami ,FL, United States, October 2011.

[65] C. Févotte, N. Bertin, and J. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis. *Neural Computation*, 21(3) :793–830, 2009.

[66] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9) :2421–2456, 2011.

[67] D. FitzGerald. User assisted separation using tensor factorisations. In *Proc. 20th European Signal Processing Conference (EUSIPCO)*, pages 2412–2416, Bucharest, Romania, Aug. 2012.

[68] D. FitzGerald, M. Cranitch, and E. Coyle. Sound source separation using shifted non-negative tensor factorisation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages V–V, Toulouse, France, May 2006.

[69] D. Fitzgerald, M. Cranitch, and E. Coyle. Extended nonnegative tensor factorisation models for musical sound source separation. *Computational Inteligence and Neuroscience*, 2008.

[70] J. Foote and S. Uchihashi. The beat spectrum : a new approach to rhythm analysis. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 881–884, Aug 2001.

[71] J. Fritsch and M. D. Plumbley. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 888–891, Vancouver, Canada, May 2013.

[72] B. Fuentes. *L'analyse probabiliste en composantes latentes et ses adaptations aux signaux musicaux. Application à la transcription automatique de musique et à la séparation de sources.* Thèses, Télécom ParisTech, Paris, France, 2013.

[73] B. Fuentes, R. Badeau, and G. Richard. Blind harmonic adaptive decomposition applied to supervised source separation. In *Proc. 20th European Signal Processing Conference (EUSIPCO)*, pages 2654–2658, Bucharest, Romania, Aug. 2012.

[74] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel. *Source separation by score synthesis.* Ann Arbor, MI : MPublishing, University of Michigan Library, 2010.

[75] T. Gerber, M. Dutasta, L. Girin, and C. Févotte. Professionally-produced music separation guided by covers. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 85–90, Porto, Portugal, Oct. 2012.

[76] Z. Goh, K.-C. Tan, and K.-C. Tan. Postprocessing method for suppressing musical noise generated by spectral subtraction. *IEEE Transactions on Speech and Audio Processing*, 6(3) :287–292, May 1998.

[77] G. Gravier, N. Souviraà-Labastie, S. Campion, and F. Bimbot. Audio thumbnails for spoken content without transcription based on a maximum motif coverage criterion. In *Annual Conference of the International Speech Communication Association*, pages 1767–1771, Singapore, Singapore, Sep 2014.

[78] R. Gribonval. Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages III–3057–III–3060, Orlando, FL, United States, May 2002.

[79] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2) :236–243, Apr 1984.

[80] G. Grindlay and D. P. W. Ellis. Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments. *IEEE Journal of Selected Topics in Signal Processing*, 5(6) :1159–1169, 2011.

[81] J. Haitsma, T. Kalker, and J. Oostveen. Robust audio hashing for content identification. In *International Workshop on Content-Based Multimedia Indexing*, volume 4, pages 117–124, Madrid, Spain, Jun. 2001.

[82] Y. Han and C. Raphael. Informed source separation of orchestra and soloist. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 315–320, Utrecht, Netherlands, August 2010.

[83] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 327–332, 2009.

[84] R. Hennequin. *Décomposition de spectrogrammes musicaux informée par des modèles de synthèse spectrale. Modélisation des variations temporelles dans les éléments sonores.* Thèses, Télécom ParisTech, Nov. 2011.

[85] R. Hennequin, J. J. Burred, S. Maller, and P. Leveau. Speech-guided source separation using a pitch-adaptive guide signal model. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6672–6676, Florence, Italy, May 2014.

[86] R. Hennequin, B. David, and R. Badeau. Beta-divergence as a subclass of Bregman divergence. *IEEE Signal Processing Letters*, 18(2) :83–86, 2011.

[87] R. Hennequin, B. David, and R. Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 45–48, Prague, Czech Republic, 2011.

[88] C. Herley. Argos : Automatically extracting repeating objects from multimedia streams. (MSR-TR-2004-02) :115–129, February 2006. IEEE Transactions On Multimedia.

[89] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties. MPEG-H Audio - The New Standard for Universal Spatial/3D Audio Coding. *Journal of the Audio Engineering Society*, 62(12) :821–830, 2015.

[90] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka. Underdetermined blind separation and tracking of moving sources based on doa-hmm. In *Proc. IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3191–3195, Florence, Italy, May 2014.

[91] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6) :82–97, Nov 2012.

[92] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7) :1527–1554, 2006.

[93] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5 :1457–1469, Dec. 2004.

[94] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. Okuno. Simultaneous processing of sound source separation and musical instrument identification using bayesian spectral modeling. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3816–3819, Prague, Czech Republic, May 2011.

[95] C. Joder and B. Schuller. Score-informed leading voice separation from monaural audio. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 277–282, Porto, Portugal, Oct 2012.

[96] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals : demixing n sources from 2 mixtures. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 2985–2988 vol.5, Istanbul, Turkey, 2000.

[97] E. Keogh and A. Ratanamahatana. Everything you know about dynamic time warping is wrong. *3rd Workshop on Mining Temporal and Sequential Data, in conjunction with 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD-2004)*, 2004.

[98] J. Keshet, D. Grangier, and S. Bengio. Discriminative keyword spotting. *Speech Communication*, 51(4) :317 – 329, 2009.

[99] E. Kijak, G. Gravier, L. Oisel, and P. Gros. Structuration multimodale d'une vidéo de tennis par modèles de markov cachés. In *19ème Colloque sur le traitement du signal et des images*, France, 2003.

[100] C. Kim and R. Stern. Power-normalized cepstral coefficients (pncc) for robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4101–4104, Kyoto, Japan, March 2012.

[101] M. Kim and P. Smaragdis. Collaborative audio enhancement using probabilistic latent component sharing. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 896–900, Vancouver, Canada, May 2013.

[102] A. Klapuri and M. Davy. *Signal processing methods for music transcription*. Springer Science & Business Media, 2007.

[103] C. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4) :320–327, Aug. 1976.

[104] M. Lagrange, A. Ozerov, and E. Vincent. Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, Porto, Portugal, Oct. 2012.

[105] L. Le Magoarou, A. Ozerov, and N. Duong. Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization. *Journal of Signal Processing Systems*, 79(2) :117–131, 2015.

[106] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong. Text-informed audio source separation using nonnegative matrix partial co-factorization. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2013)*, pages 1–6, Southampton, United Kingdom, Sept. 2013.

[107] D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401 :788–791, 1999.

[108] A. Lefevre, F. Bach, and C. Févotte. Semi-supervised nmf with time-frequency annotations for single-channel source separation. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 115–120, 2012.

[109] P. Leveau, S. Maller, J. Burred, and X. Jaureguiberry. Convolutive common audio signal extraction. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 165–168, New Paltz, NY, United States, Oct 2011.

[110] A. Liutkus. *Gaussian processes for source separation and posterior source coding.* Thèses, Télécom ParisTech, Nov. 2012.

[111] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard. An overview of informed audio source separation. In *Proc. International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4, Paris, France, Jul. 2013.

[112] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet. Kernel additive models for source separation. *IEEE Transactions on Signal Processing*, 62(16) :4298–4310, Aug 2014.

[113] A. Liutkus and P. Leveau. Separation of music+effects sound track from several international versions of the same movie. In *Proc. 128th Audio Engineering Society (AES) Convention*, London, United Kingdom, May 2010.

[114] L. Lu and A. Hanjalic. Audio content discovery : An unsupervised approach. In A. Divakaran, editor, *Multimedia Content Analysis*, Signals and Communication Technology, pages 1–39. Springer US, 2009.

[115] S. Makino, T.-W. Lee, and H. Sawada. *Blind speech separation.* Springer, 2007.

[116] M. Moussallam. *Représentations redondantes et hiérarchiques pour l'archivage et la compression de scènes sonores.* Thèses, Télécom ParisTech, 2012.

[117] M. Moussallam, G. Richard, and L. Daudet. Audio source separation informed by redundancy with greedy multiscale decompositions. In *Proc. 20th European Signal Processing Conference (EUSIPCO)*, pages 2644–2648, Bucarest, Romania, Aug. 2012.

[118] M. Müller. *Information Retrieval for Music and Motion*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[119] M. Muller and F. Kurth. Enhancing similarity matrices for music audio analysis. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 437–440, Toulouse, France, May 2006.

[120] A. Muscariello. *Variability tolerant discovery of arbitrary repeating patterns in audio data*. Thèses, Université Rennes 1, Jan. 2011.

[121] A. Muscariello, G. Gravier, and F. Bimbot. Audio keyword extraction by unsupervised word discovery. In *Annual Conference of the International Speech Communication Association*, pages 2843–2846, Brighton, United Kingdom, Sept. 2009.

[122] A. Muscariello, G. Gravier, and F. Bimbot. Zero-resource audio-only spoken term detection based on a combination of template matching techniques. In *Annual Conference of the International Speech Communication Association*, pages 921–924, Florence, Italy, Aug. 2011. spoken term detection, template matching, unsupervised learning, posterior features.

[123] A. Muscariello, G. Gravier, and F. Bimbot. Unsupervised motif acquisition in speech via seeded discovery and template matching combination. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7) :2031–2044, Sept. 2012.

[124] C. S. Myers and L. R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected-word recognition. *Bell System Technical Journal*, 60(7) :1389–1409, 1981.

[125] G. J. Mysore. *A Non-negative Framework for Joint Modeling of Spectral Structure and Temporal Dynamics in Sound Mixtures*. Thèses, Stanford University, 2010.

[126] G. J. Mysore and P. Smaragdis. Relative pitch estimation of multiple instruments. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 313–316, Taipei, Taiwan, 2009.

[127] M. Nakano, J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama. Infinite-state spectrum model for music signal analysis. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1972–1975, Prague, Czech Republic, May 2011.

[128] P. D. O'grady, B. A. Pearlmutter, and S. T. Rickard. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, 15 :18–33, 2005.

[129] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 139–144, Philadelphia, PA, United States, September 2008.

[130] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3) :550–563, March 2010.

[131] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 257–260, Prague, Czech Republic, May 2011.

[132] A. Ozerov, C. Févotte, and M. Charbit. Factorial scaled hidden markov model for polyphonic audio representation and source separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 121–124, New Paltz, NY, United States, Oct 2009.

[133] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Coding-based informed source separation : nonnegative tensor factorization approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(8) :1699–1712, August 2013.

[134] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of Bayesian models for single channel source separation and its application to voice / music separation in popular songs. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5) :1564–1578, 2007.

[135] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot. One microphone singing voice separation using source-adapted models. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 90–93, Oct 2005.

[136] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4) :1118 – 1133, May 2012.

[137] M. Parvaix. *Watermarking-based informed audio source separation for linear instantaneous stationary mixtures.* Thèses, Institut National Polytechnique de Grenoble - INPG, 2010.

[138] M. Parvaix and L. Girin. Informed source separation of linear instantaneous underdetermined audio mixtures by source index embedding. *IEEE Transactions on Audio, Speech and Language Processing*, 19(6) :1721–1733, Aug 2011.

[139] M. Parvaix, L. Girin, and J. Brossier. A watermarking-based method for informed source separation of audio signals with a single sensor. *IEEE Transactions on Audio, Speech and Language Processing*, 18(6) :1464–1475, Aug 2010.

[140] K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. Version 20121115.

[141] J. Pinquier. *Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle.* Thèses, Université Paul Sabatier-Toulouse III, 2004.

[142] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies. Sparse representations in audio and music : from coding to source separation. *Proceedings of the IEEE*, 98(6) :995–1005, June 2010.

[143] L. R. Rabiner and B.-H. Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1) :4–16, 1986.

[144] Z. Rafii and B. Pardo. A simple music/voice separation method based on the extraction of the repeating musical structure. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 221–224, Prague, Czech Republic, May 2011.

[145] Z. Rafii and B. Pardo. Music/voice separation using the similarity matrix. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, Porto, Portugal, October 8-12 2012.

[146] Z. Rafii and B. Pardo. REpeating Pattern Extraction Technique (REPET) : A Simple Method for Music/Voice Separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1) :71–82, January 2013.

[147] D. Ribas, E. Vincent, and J. R. Calvo. Uncertainty propagation for noise robust speaker recognition : the case of NIST-SRE. In *Interspeech 2015*, Dresden, Germany, Sept. 2015.

[148] J.-M. Rietsch, M.-A. Chabin, and É. Caprioli. Dématérialisation et archivage électronique. *Paris : Dunod*, 2006.

[149] N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. In *Proc. International Joint Conference on Neural Networks*, volume 4, pages 2861–2866 vol.4, 2001.

[150] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada, and S. Makino. Speech enhancement with ad-hoc microphone array using single source activity. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–6, Oct 2013.

[151] Y. Salaün, E. Vincent, N. Bertin, N. Souviraà-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot. The flexible audio source separation toolbox version 2.0. In *Show & Tell IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[152] M. N. Schmidt and H. Laurberg. Nonnegative matrix factorization with gaussian process priors. *Intell. Neuroscience*, 2008 :3 :1–3 :10, Jan. 2008.

[153] M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic latent variable models as nonnegative factorizations. *Computational intelligence and neuroscience*, 2008.

[154] U. Simsekli, Y. K. Yilmaz, and A. T. Cemgil. Score guided audio restoration via generalised coupled tensor factorisation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5369–5372, Kyoto, Japan, Mar. 2012.

[155] P. Smaragdis. Relative-pitch tracking of multiple arbitrary sounds. *The Journal of the Acoustical Society of America*, 125(5) :3406–3413, 2009.

[156] P. Smaragdis and J. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180, New Paltz, NY, United States, Oct 2003.

[157] P. Smaragdis and G. J. Mysore. Separation by humming : User-guided sound extraction from monophonic mixtures. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 69 – 72, New Paltz, NY, United States, Oct. 2009.

[158] P. Smaragdis, B. Raj, and M. Shashanka. Sparse and shift-invariant feature extraction from non-negative data. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2069–2072, Las Vegas, NV, United States, March 2008.

[159] N. Souviraà-Labastie, L. Catanese, G. Gravier, and F. Bimbot. The MODIS software for word like motif discovery and its use for zero resource audio summarization. Technical Report RT-0439, July 2013.

[160] N. Souviraà-Labastie, A. Olivero, E. Vincent, and F. Bimbot. Audio source separation using multiple deformed references. In *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, pages 311–315, Lisboa, Portugal, Sept. 2014.

[161] N. Souviraà-Labastie, A. Olivero, E. Vincent, and F. Bimbot. Multi-channel audio source separation using multiple deformed references. *IEEE Transactions on Audio, Speech and Language Processing*, 23(11) :1775–1787, June 2015.

[162] N. Souviraà-Labastie, E. Vincent, and F. Bimbot. Music separation guided by cover tracks : designing the joint NMF model. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 484–488, Brisbane, Queensland, Australia, Apr 2015.

[163] P. Sprechmann, A. Bronstein, J.-M. Morel, and G. Sapiro. Audio restoration from multiple copies. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 878–882, Vancouver, Canada, May 2013.

[164] I. Szöke, P. Schwarz, P. Matêjka, and M. Karafiàt. Comparison of keyword spotting approaches for informal continuous speech. In *Proc. Eurospeech*, pages 633–636, 2005.

[165] D. Tran, E. Vincent, and D. Jouvet. Fusion of Multiple Uncertainty Estimators and Propagators for Noise Robust ASR. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[166] D. Tran, E. Vincent, and D. Jouvet. Discriminative uncertainty estimation for noise robust ASR. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5512 – 5516, Brisbane, Queensland, Australia, Apr. 2015.

[167] D. Tran, E. Vincent, and D. Jouvet. Nonparametric uncertainty estimation and propagation for noise robust ASR. TASLP, Jan. 2015.

[168] B. Van Veen and K. Buckley. Beamforming : a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2) :4–24, April 1988.

[169] E. Vincent. *Instrument models for source separation and transcription of music recordings*. Thèses, Université Pierre et Marie Curie - Paris VI, Dec. 2004.

[170] E. Vincent. Complex nonconvex lp norm minimization for underdetermined source separation. In *Proc. 7th International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, pages 430–437, London, United Kingdom, Sept. 2007.

[171] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni. The second 'chime' speech separation and recognition challenge : An overview of challenge systems and outcomes. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 162–167, Olomouc, Czech Republic, Dec. 2013.

[172] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3) :528–537, March 2010.

[173] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot. From blind to guided audio source separation : How models and side information can improve the separation of sound. *IEEE Signal Processing Magazine*, 31(3) :107–115, May 2014.

[174] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4) :1462–1469, Jul. 2006.

[175] E. Vincent, G. Jafari, Maria, A. Abdallah, Samer, D. Plumbley, Mark, and E. Davies, Mike. Probabilistic modeling paradigms for audio source separation. In W. Wang, editor, *Machine Audition : Principles, Algorithms and Systems*, pages 162–185. IGI Global, 2010.

[176] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca. First stereo audio source separation evaluation campaign : data, algorithms and results. In *Proc. 7th International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, pages 552–559, London, United Kingdom, Sept. 2007.

[177] T. Virtanen. Algorithm for the separation of harmonic sounds with time-frequency smoothness constraint. In *Proc. International Conference on Digital Audio Effects (DAFx)*, pages 35–40, London, United Kingdom, 2003.

[178] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3) :1066–1074, March 2007.

[179] A. L. Wang. An industrial-strength audio search algorithm. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 7–13, Washington, D.C., United States, Oct 2003.

[180] Q. Wang, W. Woo, and S. Dlay. Informed single-channel speech separation using hmm-gmm user-generated exemplar source. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(12) :2087–2100, Dec 2014.

[181] N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press Cambridge, MA, 1949.

[182] M. Wölfel and J. McDonough. *Distant Speech Recognition*. John Wiley & Sons, Ltd, 2009.

[183] J. Woodruff, B. Pardo, and R. B. Dannenberg. Remixing stereo music with score-informed source separation. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 314–319, Victoria (BC), Canada, October 8-12 2006.

[184] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7) :1830–1847, July 2004.